Author:
Cornish, Rosie

Title:
Using linked health and administrative data to reduce bias due to missing data and measurement error in observational research

Author:
Cornish, Rosie

Title:
Using linked health and administrative data to reduce bias due to missing data and measurement error in observational research

Using linked health and administrative data to reduce bias due to missing data and measurement error in observational research

Rosaleen Peggy Cornish

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Medicine

Population Health Sciences, Bristol Medical School
February 4th, 2019

Word count: 48,676

Abstract

Missing data and measurement error are common problems in epidemiological studies. Missing data will lead to a loss of power and can result in bias. A complete case analysis, which uses only observations with fully observed data, will generally produce a biased estimate of the exposure-outcome association if the missingness mechanism depends on the outcome of interest. Misclassification – measurement error in a categorical variable – will always bias exposure-outcome estimates.

I use data from the Avon Longitudinal Study of Parents and Children to examine the impact of missingness and misclassification on exposure-outcome estimates by studying three epidemiological questions. I use proxies obtained via linkage (i) to examine the missing data mechanism; (ii) as auxiliary variables in inverse probability weighting (IPW), multiple imputation (MI) and full information maximum likelihood (FIML) models; and (iii) to correct for misclassification. I use simulations to evaluate bias and efficiency of these methods under a range of conditions.

I show that linked proxies can be used to establish a set of plausible missingness mechanisms and thus help identify an appropriate analysis strategy. Through simulations I demonstrate that, when the complete case analysis is biased, inclusion of proxies in MI (and FIML for a continuous outcome) will lead to reductions in bias and increases in efficiency provided the proxies are reasonably well correlated with the missing study variable. IPW may not always reduce bias and will lead to reduced precision if the proxies are also incomplete. Further, I find that MI provides a flexible way to simultaneously address missing data and misclassification and show that bias due to misclassification (in a binary exposure) is reduced even when the gold standard is missing not at random.

I provide guidance on how to approach missing data and misclassification problems when proxies are available through linkage to external datasets.

Acknowledgements

I am extremely grateful to my supervisors, John Macleod and Kate Tilling, for all the invaluable advice, guidance and support they have given me throughout my PhD.

I would like to thank Rachael Hughes for her help and advice on setting up simulation studies and David Carslake for helping me use the university's high performance computing facilities to carry out these simulations.

Some of the work presented in this thesis was done in collaboration with others and has been published. I would like to thank the co-authors of these papers – Ann John, James Carpenter, Andy Boyd, and Amy Davies (as well as my supervisors, already mentioned) – for their valuable input.

I would also like to thank the UK Medical Research Council for providing funding for this PhD through one of their Population Health Scientist pre-doctoral fellowships.

Finally, I would also like to acknowledge all the families who took part in the Avon Longitudinal Study of Parents and Children (ALSPAC), the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that is has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: …….. DATE:

Contents

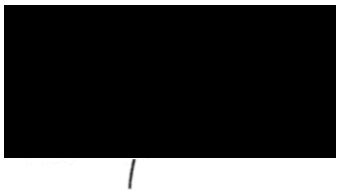
List of tables

List of figures

Publications resulting from this work

1. Cornish RP, Tilling K, Boyd A, Davies A, Macloed J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *International Journal of Epidemiology* 2015 44(3):937-45. https://doi.org/10.1093/ije.dyv035.

   Authors' contributions:
   RPC, KT and JM conceived the study. RPC conducted the statistical analyses and wrote the first draft of the manuscript. RPC, KT, AD and JM contributed to the interpretation of the results. AB designed and established the linkage and data management processes. All authors contributed to the drafting and revising of the manuscript and all read and approved the final version.

   Signed 

   Rosie Cornish (first author)

   John Macleod (last author)

2. Cornish RP, John A, Boyd A, Tilling K, Macleod J. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. *BMJ Open* 2016 6:e013167. https://doi.org/10.1136/bmjopen-2016-013167

   Published authors' contributions:
   RPC and AJ conceived the study. RC conducted the statistical analyses and wrote the first draft of the manuscript. JM, KT and AJ contributed to the interpretation of the results. AB designed and established the linkage and data management processes. All authors contributed to the design of the study and the drafting of the manuscript. All authors have read and approved the final version.

3. Cornish RP, Macleod J, Carpenter JR, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. *Emerging Themes in Epidemiology* 2017 14:14. https://doi.org/10.1186/s12982-017-0068-0

   Published authors' contributions:
   RC and KT conceived the study. RC conducted the statistical analyses and wrote the first draft of the manuscript. JC and KT contributed to the interpretation of the results. All authors contributed to the design of the study and the drafting and revising of the manuscript. All authors have read and approved the final manuscript.

List of abbreviations

ALF	Anonymised linking field
ALSPAC	Avon Longitudinal Study of Parents and Children
BMI	Body mass index
BNSSSG	Bristol, North Somerset, Somerset and South Gloucestershire
CCEI	Crown Crisp Experiental Index
CD	Compact disc
CI	Confidence interval
CIS-R	Revised Clinical Interview Schedule
CMD	Common mental disorders
DAG	Directed acyclic graph
EHC	Education, health and care
EM	Expectation-maximisation
EMIS	Egton Medical Information Systems
EPDS	Edinburgh Postnatal Depression Scale
EU	European Union
FCS	Fully conditional specification
FIML	Full information maximum likelihood
FMI	Fraction of missing information
GCSE	General Certificate of Secondary Education
GP	General Practitioner
HRA CAG	Health Research Authority Confidentiality Advisory Group
HSCIC	Health and Social Care Information Centre
ICD-10	International Classification of Diseases, 10th revision
IPW	Inverse probability weighting
IQR	Interquartile range
KS2, KS3, KS4	Key Stage 2, Key Stage 3, Key Stage 4
LMIC	Low and middle income countries
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MI	Multiple imputation
MICE	Multiple imputation using chained equations
MNAR	Missing not at random
MCSE	Monte Carlo standard error
NHS IC	National Health Service Information Centre
NPD	National Pupil Database
NPV	Negative predictive value
NWIS	NHS Wales Information Service
PBA	Probabilistic bias analysis
PEARL	Project to Enhance ALSPAC through Record Linkage
PPV	Positive predictive value
SAIL	Secure Anonymised Information Linkage
SDQ	Strength and Difficulties Questionnaire
SEN	Special educational needs
SWCSU	South West Commissioning Support Unit
WASI	Wechsler Abbreviated Scale of Intelligence

Chapter 1 Introduction

In this chapter I will provide some background in terms of the key issues being addressed in this thesis. I will also outline the structure of the thesis and its aims and objectives.

1.1 Background: sources of error in epidemiological studies

Epidemiological studies provide a vital source of evidence about the causes of ill-health as well as protective factors and potential treatments. However, all epidemiological studies have limitations. In particular, error can be introduced in a number of different ways. These sources of error can be classified into two broad categories – error due to selection processes and error due to measurement (Hennekens and Buring 1987, Webb and Bain 2011). Both types of error can lead to bias. Selection bias can be introduced in the recruitment phase of a study if individuals who are selected or who agree to take part in a study differ systematically from those who do not and the reasons for this relate to the factors being investigated. However, in this thesis I will be focussing on another potential source of selection bias – missing data arising through loss to follow-up or non-response – which, again, can lead to bias if there are systematic differences between those who remain in the study and those who do not. Measurement error is essentially the difference between the observed value of a particular measure and its true value; this can either be random or systematic error (Webb and Bain 2011). Some degree of measurement error will almost always be present in epidemiology. In some situations this may simply result in a loss of precision; in others it will lead to bias.

In this thesis, I will be examining these issues in the context of observational studies – studies in which the researcher "observes" and records relevant behaviour and outcomes through questionnaires, interviews, or clinical measurements but does not intervene in any way (for example, by administering a treatment). Although most of the work will be generalisable to randomised controlled trials (RCTs), some of the issues – for example, missingness in an exposure variable – will not be applicable in this context.

Clearly, a loss of precision (or power) is a problem in epidemiology as low power may result in inconclusive results. Likewise, understanding the impact of missing data and measurement error in terms of bias is very important because bias could lead to incorrect conclusions being drawn from a study which, in turn, could impact on healthcare and policy decisions.

1.1.1 Missing data

Missing data arising through loss-to-follow-up or non-response inevitably results in a loss of statistical power. Whether or not bias is introduced in a particular analysis depends on the analytical model used and the process that caused the data to be missing - the missing data (or missingness) mechanism. In 1976 Rubin (Rubin 1976) suggested a classification of missing data mechanisms, which has now been widely adopted. This is outlined in Box 1-1.

Box 1-1 Missing data mechanisms

Missing completely at random (MCAR)

A particular observation is said to be missing completely at random (MCAR) if the probability of it being missing does not depend on either observed or unobserved factors or, in other words, the missing observations in a dataset are simply a random subset of the whole set of observations (observed + missing). For example, if a particular piece of electronic equipment failed to work occasionally, the measurements taken using that piece of equipment would be MCAR.

Missing at random (MAR)

An observation is missing at random (MAR) if the probability that it is missing only depends on the observed data – in other words, on other factors that have been measured. For example, in a longitudinal study of changes in smoking habits among adults aged 25-50 information is collected at baseline and then five years later. If everyone completed the baseline questionnaire (which collected information on age and education as well as smoking) but those aged 25-30 and those who with lower levels of education were less likely to complete the follow-up questionnaire, then smoking status at five years follow-up would be missing at random, conditional on age and education level. Or, put another way, once age and education level have been taken into account, everyone has the same probability of having missing follow-up data on smoking.

Missing not at random (MNAR)

Finally, an observation is missing not at random (MNAR) if the probability that it is missing depends on its own (unobserved) value, even after taking account of all the observed data (i.e. all the factors that have been measured). So, if – in the above example – those who remained heavy smokers were less likely to complete the follow-up questionnaire and this propensity to remain a heavy smoker could not be explained by measured factors, then smoking status at follow-up would be missing not at random. The data would also be MNAR if there were an unmeasured factor associated with smoking status that affected response to the follow-up questionnaire.

1.1.1.1 Implications of the missing data mechanism

A common strategy used in the presence of missing data is to carry out a complete case (or complete records) analysis (Eekhout et al. 2012, Sterne et al. 2009, White et al. 2011), which only includes those observations that contain complete information

on all variables included in a particular analysis; most statistical packages do this automatically. If the data are MCAR then such an analysis will produce unbiased results (Carpenter and Kenward 2013), although with loss of efficiency. If the data are MAR or MNAR then whether or not a complete case analysis produces biased estimates of an exposure-outcome association essentially depends on whether the missingness mechanism is related to the outcome in the analysis model of interest; this is summarised in Table 1-1.

Unfortunately, using only the observed data, it is generally not possible to tell whether data are MAR or MNAR because, by definition, the values of the missing observations are unknown; thus sensitivity analyses are recommended in order to ascertain the impact of particular assumptions on the results (Carpenter and Kenward 2013).

Table 1-1: Implications of the missing data mechanism on estimates of the exposure-outcome association from a complete case analysis: multiple linear regression and multiple logistic regression

Missingness depends on:	Linear regression	Logistic regression
Neither outcome, exposure or covariates[2]	Unbiased[1]	Unbiased[1]
Exposure and/or covariates	Unbiased[1]	Unbiased[1]
Outcome	Biased[1]	Unbiased[1]
Outcome & covariates but not exposure	Biased[1]	Unbiased[1]
Outcome, exposure and possibly covariates	Biased[1]	Biased[1,3]

1. Under the assumption that the covariates are measured and included in the analysis model and the analysis model is correct (i.e. covariates are measured and, in the absence of missing data, the estimate would be unbiased)
2. Covariates = all covariates other than the exposure of interest
3. This will generally be biased. But, if there is no multiplicative interaction between the exposure and outcome in their effect on the probability of missingness, a complete case logistic regression will be unbiased (Bartlett et al. 2015). If missingness also depends on covariates, a similar condition needs to hold for the estimate to be unbiased (see Bartlett et al. 2015).

1.1.2 Measurement error

Measurement error arises as a result of either random error or systematic error (or both) and means that the observed measurements are different from their true value. Random error is the variability in repeated measurements of the same attribute resulting from a combination of the short-term variation in the characteristic being measured and the precision of the equipment used to take the measurements. Some types of measure are subject to greater random error than others. For example, a person's blood pressure varies throughout the day but a person's height (as an adult) will remain relatively stable. Similarly, some measuring devices will be more precise than others. So, random error causes measurements to vary but, on average, they will be equal to the true value.

In contrast, systematic error results in the observed measurements being consistently different from their true value (Hutcheon et al. 2010). As such, the measurements are not, on average, equal to the true value. Systematic error can be introduced in a number of ways. Individuals taking part in an observational study are often asked to answer questions about their characteristics or behaviours; these may be subject to error, perhaps because – for complex reasons which may vary from person to person – they have a tendency to under-report (for example, alcohol intake) or over-report (for example, amount of physical exercise). Other sources of systematic measurement error include observer bias, whereby there may be differences in the way in which a given attribute is measured for different groups of individuals; and the validity of the instrument being used to measure a particular variable - in other words, whether the instrument actually measures what it is supposed to be measuring (Hennekens and Buring 1987, Webb and Bain 2011).

In epidemiology, where variables are often categorical, measurement error is analogous to misclassification – wrongly classifying subjects in terms of their exposure, outcome or covariate status. In this thesis I will focus on misclassification (rather than measurement error in continuous variables). Misclassification can either be differential or non-differential. If misclassification is non-differential, all individuals

in a study have the same probability of misclassification, regardless of the true value of the variable of interest (Hennekens and Buring 1987, Webb and Bain 2011). Conversely, if misclassification is differential, some subjects are more likely to be misclassified than others (Hennekens and Buring 1987). For example, very heavy smokers might be more likely to be misclassified (for example, as moderate smokers) than light smokers or non-smokers.

One key difference between measurement error in a continuous variable and misclassification is that random measurement error in a continuous variable is unrelated to (independent of) its true value whereas misclassification (even if random), is always related to the true value (so, for example, with a binary exposure or outcome, misclassified individuals who are truly unexposed or without disease will always be (mis)classified as exposed/diseased, and vice versa). As a result, misclassification will always result in bias.

1.1.2.1 Effect of misclassification

Exactly what happens to an effect estimate in the presence of misclassification depends on several factors. These include whether or not the variable is binary, whether it is the outcome, exposure, or a confounder in the specific analysis being conducted and whether the misclassification is non-differential or differential. This is summarised in Table 1-2. If confounders are misclassified, this will lead to incomplete adjustment for confounding and thus any estimate of association between exposure and outcome will be biased (Greenland 1980). If a binary exposure or outcome is subject to non-differential misclassification, then this generally produces bias towards the null (Copeland et al. 1977). This is also true if both the exposure and outcome are subject to non-differential misclassification, as long as the misclassification in the exposure and the outcome are independent of each other (Kristensen 1992). However, it should be noted that bias is a measurement of the average effect on the estimate across repeated studies and, contrary to what is often claimed, does not mean that the observed association in a study will itself be an under-estimate of the true association (Jurek et al. 2005). In fact, although the observed association is more

likely to be an under-estimate, it could be the opposite – due to sampling variability (Jurek et al. 2005). When the exposure variable has more than two categories then the bias can be in either direction (Dosemeci et al. 1990). Similarly, differential misclassification can also bias results in either direction, depending on the nature of this misclassification (Copeland et al. 1977).

Table 1-2: Effect of misclassification on estimates of the exposure-outcome association

	Non-differential	Differential
Confounder	Bias due to residual confounding; bias can be in either direction	
Binary outcome or exposure	Bias towards the null	Bias in either direction
Outcome or exposure with >2 categories	Bias in either direction	

1.2 The role of linked data in addressing bias

Linkage to routine health or administrative datasets containing equivalent or proxy measures of variables measured in an observational study offers one way of understanding and addressing bias due to missing data and measurement error. Although the information contained in such datasets is often less detailed than data collected as part of an observational study, they are generally population-based and thus don't suffer from non-response in the same way. Having said this, such datasets may suffer from missing information due to: selection processes relating to coverage; lack of information on identifiers needed to establish a link; lack of consent for linkage; or, particularly in the case of health datasets, behavioural factors associated with use of health services and factors relating to information systems (for example, how information is recorded and stored). In addition, the data are typically – although not always – more objectively measured and thus potentially less subject to information bias. Previous studies have used linkage to alternative sources of data to address bias due to either measurement error or missing data; these are reviewed in Chapter 2. However, many of these studies have simply used the linked variable in place of the original study variable and/or focused on the impact of missing data or

misclassification on estimates of prevalence, incidence or the mean value of a given outcome variable, rather than examining bias in estimates of exposure-outcome associations.

In this thesis I examine both missingness and misclassification in exposure as well as outcome variables and explore the potential for addressing these sources of bias using linked data. In terms of missing data, I specifically focus on exposures and outcomes that, a priori, I expect to be MNAR. I analyse data from the Avon Longitudinal Study of Parents and Children (ALSPAC). Further, using simulation studies based on three exemplars, I examine different scenarios – both in terms of the extent of missingness and in terms of the type of analysis being undertaken – and compare results obtained using different methods that attempt to account for missing data. I examine misclassification in the presence of a gold (reference) standard. By systematically investigating a wide range of different scenarios, I demonstrate the conditions under which linkage to external sources of data is likely to be beneficial in terms of leading to reductions in bias and increases in efficiency.

The simulations are important in this context because, in a simulation study, the "truth" (usually the true value of the parameter – or parameters – of interest) is known because the data have been generated (via a known process). Since the truth is known, different methods can be evaluated and compared because performance measures (such as bias) can be calculated (Morris et al. 2017).

1.3 Aims and objectives

The overarching aim of this thesis is to examine how linked health and administrative data can be used to both understand and reduce bias due to missing data and measurement error in prospective cohort studies (as stated above, I will only focus on misclassification not measurement error in continuous variables), using ALSPAC as the exemplar setting. This aim is addressed using simulation studies and by examining three questions of epidemiological importance. These are presented in detail in Chapter 2 (Section 2.5) but are outlined briefly here:

1. Is breastfeeding associated with IQ at age 15?

Approximately 12,500 young people have data on breastfeeding but only 39% (just over 4,900) of these have data on IQ at age 15. Linkage to education data (GCSE results and other information) is used to examine the missingness mechanism for IQ, and in imputation of the missing values.

2. Is maternal smoking in pregnancy associated with depression at age 17-18?

The mothers of just over 11,000 children gave information about smoking during pregnancy. In contrast, only around 4,500 (~40%) young people came to the clinic at age 17-18 and completed the computerised depression questionnaire. As with smoking, linkage to relevant data held within GP records is used to look at the objectives below in relation to this outcome.

3. Is teenage smoking associated with educational attainment at age 16?

Children were asked in detail about smoking yearly between the ages of 12 and 15 years old. GCSE results from linked education data are available for just under 12,000 subjects but only 56% of these completed at least one of the above questionnaires about smoking (<40% at age 15 years). Data on smoking from the young people's GP records is used to examine missing data patterns in ALSPAC-measured smoking and to investigate misclassification.

1.3.1 Objectives

Two of the objectives relate to missing data and two to misclassification.

Missing data objectives:

1. To use linked health and administrative data to examine patterns of missing data and to model missingness mechanisms in a longitudinal study (ALSPAC), focussing in particular on outcomes and exposures that are likely to be MNAR.

2. To incorporate linked health and administrative data as auxiliary variables in multiple imputation and other models to explore bias in estimates of exposure-outcome associations introduced by missing data in exposures or outcomes.

Misclassification objectives:

3. To compare self-reported smoking data in ALSPAC and smoking recorded in linked electronic primary care records (GP data) to a gold standard measure of smoking in order to investigate misclassification in self-reported and GP-recorded smoking and, in particular, to identify whether these are subject to differential or non-differential misclassification.

4. To explore methods for using both linked and self-reported data to minimise the impact of misclassification on analyses in observational studies.

1.4 Structure of the thesis

This thesis has nine chapters. Chapters 1 to 3 are introductory chapters and the final chapter summarises the main findings, discusses the implications of these and draws some conclusions. The central five chapters (Chapters 4 to 8) cover the methods and results. Chapters 4 to 7 cover missing data and Chapter 8 describes the methods and results relating to misclassification. Further details are given below.

Chapters 1 to 3

The aim of Chapter 1 is to introduce the thesis in terms of its structure and aims and to give a brief background regarding the key sources of error in observational

research. Chapters 2 and 3 provide further background: in Chapter 2 I give details of the statistical methods commonly used to take account of missing data and misclassification and review how – in the literature – linked data have been used to address these sources of bias; in Chapter 3 I describe ALSPAC and the data used for this research, including (where appropriate) code-based algorithms previously used to define the exposures and outcomes used in this thesis.

Chapters 4 to 7

These chapters all address missing data. In Chapter 4 I describe work I did in relation to examining predictors of participation in ALSPAC as well as predictors of inclusion in the GP and education datasets. I also discuss the implications of this for the exemplars used in this thesis. Chapters 5 to 7 describe the methods and results in the case of (i) missingness in a continuous outcome, (ii) missingness in a binary outcome, and (iii) missingness in a categorical exposure (respectively). In each case I describe the relevant exemplar study and then present the simulation studies I carried out to examine the impact of missing data and of having linked proxies for the missing variables.

Chapter 8

This chapter focuses on misclassification. Specifically, I consider misclassification in a binary exposure when a gold standard measure is available for a subset of the individuals; this is based on Exemplar 3 (teenage smoking and educational attainment).

Chapter 9

The final chapter contains a summary of the rationale, objectives and the main findings of this thesis. In addition, I discuss the strengths and limitations of the research and draw some overall conclusions. I also discuss the implications of the work, both in terms of current practice and in terms of future research.

Chapter 2 Background

This chapter is divided into five sections. The first two sections provide background on the statistical methods I will use in this thesis. Section one covers methods used to analyse incomplete data and section two covers methods used to take account of misclassification. The third section summarises previous studies that have used linked data to address bias due to missing data or measurement error. Sections four and five describe the current evidence relating to (i) participation in cohort studies, including ALSPAC, and (ii) the exemplar questions being considered (respectively).

2.1 Analysing incomplete data

There are several different methods that can be used to minimise bias and loss of efficiency due to missing data. In this chapter I will describe only those methods used in this thesis. These include a complete case analysis, multiple imputation using chained equations, inverse probability weighting and full information maximum likelihood. In this thesis I am only focussing on bias in estimates of exposure-outcome associations. Thus, when I discuss bias I will referring only to bias in the coefficient(s) of X (the exposure) when carrying out a linear regression of Y (the outcome) – or, in the case of a binary outcome, logistic regression – on X, with or without additional covariates.

2.1.1 Complete case analysis

In a complete case analysis, only observations with complete data for all the variables in the analysis of interest are included; statistical packages will do this by default. This is also sometimes referred to as listwise deletion. One key advantage of a complete case analysis is that it will produce an unbiased estimate of the exposure-outcome association if the missingness mechanism only depends on the exposure or covariates included in the model (Carpenter and Kenward 2013, Little 1992). Further, if the outcome is binary and logistic regression is used, a complete case analysis will produce an unbiased estimate of the exposure-outcome association unless there is a multiplicative interaction between the exposure and outcome with respect to the probability of missingness (Bartlett et al. 2015). (Note that if missingness also depends on covariates, a similar condition needs to hold for the estimate to be unbiased – see Bartlett et al. 2015.) Having said this, one clear disadvantage of a complete case analysis is that it will result in a loss of efficiency; the extent of this will depend on the amount of missing information.

2.1.2 Inverse probability weighting

Inverse probability weighting (IPW) is a two-stage process. The first stage involves estimating the probability (for each observation) of having complete data. This is typically done using logistic regression. In the second stage a complete case analysis is carried out in which each observation is weighted by the inverse of the fitted probabilities of having complete data (obtained from the first model) (Vansteelandt et al. 2010). As such, observations with a low probability of being complete are given greater weight than those that are more likely to be complete. For example, if missingness was only dependent on sex and males had a probability of 50% of being a complete case whereas females had a probability of 75%, then all the males with complete data would receive a weight of 2 and all females with complete data would receive a weight of 4/3.

In the context of this thesis I will be using the linked variables (i.e. those obtained from external datasets) as additional predictor variables in stage one of this process (deriving the weights). As above, where the missing study variable is MNAR, these linked variables are likely to be predictors of the missing values as well as predictors of missingness and are therefore suitable for generating weights in this context (Seaman and White 2013).

One problem that can occur in IPW is when a small number of observations receive very large weights; this may result in large standard errors of the estimate(s) of interest (Vansteelandt et al. 2010). This can happen if (a) the variables included in the (missingness) model are strongly predictive of missingness or (b) the missingness model is incorrectly specified. It is thought that the second explanation is more likely (Seaman and White 2013). In this situation one way of dealing with large weights is to truncate them. Here, a maximum weight is chosen and any weights that are larger than this maximum are set to its value (i.e. the value of the maximum specified weight) (Seaman and White 2013). Seaman and White recommend that a sensitivity analysis should be carried out, varying the value of the maximum weight.

2.1.2.1 Validity of IPW

IPW will be valid if the missingness model (the model to calculate the weights) is correctly specified, which implies that the data need to be MAR. As a consequence, IPW only works well if you have fully observed predictors of missingness or data that are monotone missing – variables v_1 to v_k are said to be monotone missing if: when v_j is missing then all v_{j+1} are also missing (for all j=1 to k) (Seaman and White 2013). This can occur in longitudinal studies when individuals drop out (and contribute no further data from this point). However, data in longitudinal studies will be non-monotone missing when individuals respond to some rounds of data collection (i.e. questionnaires, study clinics, etc) but not others and/or, within a given data collection, do not respond to all items. In this situation, standard IPW methods as described above will not be valid (Sun et al. 2018). IPW methods do exist for non-monotone missing data. Robins and Gill used a Markov randomised monotone

missingness model (Robins and Gill 1997). However, this approach has yet to be incorporated in any software and only works well if the number of incomplete predictors of missingness is small (Seaman and White 2013, Sun and Tchetgen Tchetgen 2018) and, as a result, has not been widely used to date. More recently, others have demonstrated the use of unconstrained maximum likelihood and constrained Bayesian estimation for non-monotone missing data (Sun and Tchetgen Tchetgen 2018).

2.1.3 Multiple imputation

Multiple imputation (MI) was first proposed by Rubin (Rubin 1976, Rubin 1987). It is a three-stage process and is a Bayesian-based procedure. In the first stage, a number (M) of complete datasets are created by imputing values for the missing data. The imputed values are drawn from the conditional posterior predictive distribution of the missing data (conditional on the observed data and the missingness mechanism or, under the MAR assumption, conditional only on the observed data) (Carpenter and Kenward 2013). Although it is possible to impute under a MNAR mechanism (Galimard et al. 2016, Schafer 2003), standard implementations of multiple imputation assume the data are MAR (Schafer 1999, White et al. 2011).

In the second stage, the M imputed datasets are each analysed identically using standard statistical techniques. In the final stage, the estimates obtained from each of these datasets are combined by taking the arithmetic mean to obtain an overall estimate. The standard error of this overall estimate is obtained using Rubin's rules (Rubin 1987), which takes account of the variation both within and between imputations. So, suppose we are trying to estimate a parameter β with variance V. Each of the M datasets provides an estimate of the parameter of interest, $\hat{\beta}_i$ with variance $\hat{V}_i$ for $i = 1, \dots, M$. The overall estimate of β is given by:

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_i$$

If we define

$$\hat{V}_W = \frac{1}{M}\sum_{i=1}^{M}\hat{V}_i$$

as the average within imputation variance of β and

$$\hat{V}_B = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\beta}_i - \hat{\beta})^2$$

as the between imputation variance, then the estimate of the overall variance of β is given by

$$\hat{V} = \hat{V}_W + \left(1 + \frac{1}{M}\right)\hat{V}_B$$

2.1.3.1 Multiple imputation using chained equations

Except in relatively straightforward cases in which Bayesian parametric models can be used (Schafer 1999), Markov chain Monte Carlo (MCMC) methods need to be used to generate the imputations (Schafer 1997). MCMC methods are simulation-based techniques used to draw repeated samples from the conditional posterior distribution of a parameter (or set of parameters) – given the data – in order to generate repeated estimates of this parameter. The observed distribution of these estimates is then used to approximate the distribution of the parameter (and thus approximate statistics such as the expected value of this parameter) (Hoff 2009). In the context of multiple imputation, MCMC methods are used to sample from the conditional distribution of the parameters of interest as well as the conditional distribution of the missing data (Carpenter and Kenward 2013).

When multiple imputation was first proposed, the procedure involved jointly modelling all the variables – for example, assuming a multivariate normal model (Rubin 1987, Schafer 1997). However, it is not always possible to specify a joint model for the data (van Buuren 2007). Multiple imputation using chained equations (MICE), also known as fully conditional specification (FCS), is a more flexible approach that can be used in situations where the missing data occurs in a mixture of different types

of variable (continuous and categorical) (van Buuren 2007, White et al. 2011). In MICE, rather than jointly modelling the variables, there is a separate imputation model for each of the incomplete variables. The process is as follows (Azur et al. 2011, van Buuren et al. 1999, White et al. 2011):

1. Initially all the missing values in the dataset are replaced by values obtained by simply sampling from the observed values.

2. The variables containing missing values, x_j are ordered such that x_{j-1} contains a lower proportion of missing data than x_j for all j.

3. The values that were initially missing for x_1 are re-set to missing. x_1 is then regressed on all the other variables and missing values are imputed from its posterior predictive distribution.

4. Step 3 is then repeated for each of the remaining variables $x_2$to x_k. In each case, the regression model includes individuals with observed values for the variable being imputed but uses imputed values of all other variables.

5. Once all variables have been imputed once as part of this process, this completes a cycle. This process (i.e. steps 3 and 4) is repeated for a number of cycles – typically 10 – to allow the imputed values to stabilise; this results in one imputed dataset.

6. Steps 3 to 5 are then repeated M times to produce M imputed datasets.

Because each variable is imputed separately, different regression models (linear, logistic, and so on) can be used to impute the variables with missing values. Although there is no theoretical justification for the chained equations approach unless the models would converge to an underlying joint model, research to date suggests that valid results are obtained (Hughes et al. 2014).

When Rubin first suggested multiple imputation as a method for handling missing data, he argued that only a small number of datasets (for example, 3 or 5) needed to

be imputed in order for inferences to be valid (Schafer 1997). However, since then it has been recommended that a greater number of imputations should be carried out (Carpenter and Kenward 2013, White et al. 2011). White et al. suggested that the number of imputations should be greater than or equal to the percentage of incomplete cases in order to minimise the Monte Carlo error and thus maximise the reproducibility of the results (White et al. 2011). In practice, it is now common to impute 100 datasets (Carpenter and Kenward 2013).

2.1.3.2 Auxiliary variables in multiple imputation

One of the key characteristics of MI is that the process of imputing the missing data and the subsequent analysis are two separate processes. This has both positive and negative consequences. The main drawback is that if the imputation model does not include all variables that are in the analysis model and/or does not incorporate important features of the analysis model, such as non-linear or interaction terms, then the multiply-imputed estimates will be biased (Kenward and Carpenter 2007, Tilling et al. 2016). In contrast, the key advantage of MI is that variables can be included into the imputation model that are not included in the subsequent analysis model in order to make an MAR assumption more plausible or to improve efficiency (or both). In particular, it is recommended that – in addition to all variables included in the analysis model – the imputation model should also include all variables that predict the incomplete variable(s) (Collins et al. 2001, Sterne et al. 2009, White et al. 2011). Variables that are included in the imputation model but not in the analysis model are referred to as auxiliary variables (White et al. 2011) and could include previous or future measurements of the incomplete variable. Including in the imputation model variables that are only predictive of missingness (but not predictive of the incomplete variables) will not reduce bias and may result in a loss of efficiency (Meng 1994). However, Collins et al. suggest adopting an inclusive strategy when using auxiliary variables; their research showed that the inclusion of all variables thought to be important in terms of predicting either missingness or the values of the incomplete variable(s) themselves will reduce bias and will only have a relatively small impact on efficiency (Collins et al. 2001).

In this thesis I will be using proxies for the missing study variables, obtained via linkage to external datasets, as auxiliary variables in MI models. In the situation where the missing study variable is MNAR, such auxiliary variables will be both predictors of the incomplete variable and predictors of missingness.

2.1.3.3 Validity of multiple imputation

As stated above, standard implementations of MI assume the data are MAR. In addition, for MI to be valid the imputation model must be correctly specified, both in terms of the variables included and in terms of its distributional assumptions (Carpenter and Kenward 2013, Nguyen et al. 2017). Importantly, the imputation model must be compatible with the analysis model, which means that – as mentioned above – it should include all variables from the analysis model, including non-linear and interaction terms (Bartlett et al. 2015, Tilling et al. 2016).

2.1.4 Full information maximum likelihood

As the name implies, full information maximum likelihood (FIML) estimation uses maximum likelihood methods to handle missing normally distributed data (Enders 2010). In the general multivariate case, the log likelihood function is given by: $\log \mathrm{L}\,(\boldsymbol{\theta}/\mathbf{Y}) = \sum_{i=1}^{n} \log f(\boldsymbol{Y_i}/\boldsymbol{\theta}))$, where $f(\boldsymbol{Y}/\boldsymbol{\theta})$ is the joint density function for a set of variables Y dependent on the vector of parameters $\boldsymbol{\theta}$. Each individual in a given dataset contributes $\log \mathrm{L}_i$ to the log likelihood function. In the case of complete data, this function is the same for all individuals in the dataset; in the presence of missing data, the log likelihood for an individual only contains terms relevant to the variables (and associated parameters) observed for that individual (Enders 2010). Thus, FIML uses all available data in order to find maximum likelihood values for the parameters of interest.

2.1.4.1 Auxiliary variables in FIML

Two approaches for incorporating auxiliary variables in FIML have been outlined; both use structural equation modelling (Graham 2003). The first one – the extra dependent variable model – includes each auxiliary variable as an extra dependent

variable (predicted by the same set of covariates as the outcome variable). Additionally, the residuals from this model are specified as being correlated with the residuals from the main outcome model and, if there is more than one auxiliary variable, their residuals are specified as being correlated with each other (Graham 2003). The second approach is called the saturated correlates model. In this model the auxiliary variables are specified as being correlated with all covariates or the residuals for any covariates that are predicted by latent variables and, as previously, all auxiliary variables (if there are more than one) are specified as being correlated (Graham 2003). If there are latent variables in the analysis model then the saturated correlates model has been shown to perform better overall. However, when the model only includes measured variables, the two models have been shown to give identical estimates (Graham 2003). In this thesis I use the extra dependent variable model, as I have no latent variables.

2.1.4.2 Validity of FIML

FIML will produce unbiased parameter estimates when data are MCAR or MAR, conditional on the variables in the analysis model (plus any auxiliary variables), provided that the analysis model is correctly specified. Note that these are the same conditions required for MI to be valid.

2.1.5 Comparison of FIML, MI and IPW

If the same assumptions are made regarding joint distributions of and relationships between variables, then FIML and MI are asymptotically equivalent (Schafer 2003). Although FIML can be extended to include auxiliary variables, it is easier to incorporate these into MI (Collins et al. 2001, Dong and Peng 2013); this is a key advantage of MI. Because IPW only uses data from complete cases in the analysis model, it tends to be less efficient than multiple imputation (Seaman and White 2013). Efficiency will be further compromised if the weighting model includes auxiliary variables that are not fully observed. However, MI will produce biased results if the imputation model is incorrectly specified and it has been argued that IPW might be preferable in situations when this is likely to be the case – for example,

when the distribution of covariates is likely to be quite different among those with and without missing data (Vansteelandt et al. 2010). Having said this, IPW will of course produce biased results if the missingness model is mis-specified (Seaman and White 2013).

2.2 Taking account of misclassification

If a gold standard measure is available for all individuals through linkage to an external dataset, then this gold standard measure can be used in place of the study variable. If a gold standard measure is available via an external validation study or is only available on a subset of individuals (through an internal validation study or through linkage on a subset), then several approaches are possible. These are outlined below.

2.2.1 Fixed or probabilistic bias analysis

One approach is to correct the original measures (Greenland and Kleinbaum 1983, Lash et al. 2009, Lyles et al. 2007, Marshall 1990, van Walraven 2017) using estimates of the sensitivity and specificity or positive and negative predictive values (PPV and NPV) obtained from the validation study. This method has been referred to in the literature as (quantitative) bias analysis (Lash et al. 2009). The estimates of sensitivity and specificity (or the predictive values) can either be regarded as fixed (fixed bias analysis) or having a probability distribution (probabilistic bias analysis) (Lash et al. 2009). Probabilistic – but not fixed – bias analysis can be applied to individual records in order to take account of covariates (and thus used to obtain adjusted exposure-outcome estimates that take account of misclassification) (Banack et al. 2018, Funk and Landi 2014). This process uses Monte Carlo methods and is outlined briefly below. It involves three main steps. In this summary I will assume that the exposure is misclassified; however, the process is the same for a misclassified outcome variable.

Step 1: Modelling the bias parameters

In order to carry out the correction, estimates of the positive and negative predictive values are needed. Data from the validation study are either used to estimate these directly or indirectly via estimates of the sensitivity and specificity (Marshall 1990). As a result, these methods have been referred to as the direct and indirect method. This is explained in greater detail in Box 2-1.

If exposure misclassification is assumed to be independent of outcome status and other covariates, a single estimate of the predictive values (PPV and NPV) can be calculated from the validation data. Alternatively, if misclassification is thought to depend on outcome status and/or other covariates, separate estimates can be calculated; these could be estimated from separate cross-tabulations or via logistic regression (Banack et al. 2018).

Once the PPV and NPV have been estimated – either directly or indirectly – beta distributions are used to model the probability density functions of these parameters.

Box 2-1: Indirect and direct method of modelling bias parameters in probabilistic bias analysis (Marshall 1990)

Indirect method

If the validation study is designed such that individuals are sampled on the basis of the imperfect (misclassified) exposure or outcome measure, then it is not possible to estimate the predictive values directly. In this case it is necessary to estimate the sensitivity and specificity from the validation data and estimate the predictive values from these.

Direct method

If individuals in the validation study are sampled on the basis of their true (gold standard) exposure or outcome status, it is not possible to estimate the sensitivity or specificity but predictive values can be estimated directly from the data.
If a random sample of all individuals is included in the validation study then both sets of parameters (predictive values and sensitivity and specificity) can be estimated. In this case, it has been shown that it is slightly more efficient to use the direct method.

Step 2: Correction for misclassification

If the validation study is internal, then if an individual has had the true exposure (gold standard) measured, this is left unchanged. For all other subjects, an individual's true exposure status is modelled as a Bernoulli random variable with probability given by a sampled value of the PPV (if classified as exposed according to the imperfect measure of exposure) or 1-NPV (if classified as unexposed according to the imperfect measure). This gives a single dataset with exposure status corrected for misclassification.

Step 3: Estimate the parameter of interest

The parameter of interest is estimated in this corrected dataset (in this thesis, I am interested in corrected estimates of the exposure-outcome association, so the parameter will be a regression coefficient or log odds ratio but in other situations the prevalence of exposure might be the parameter of interest).

Steps 2 and 3 are then repeated many times and the estimates obtained in each dataset are saved. The median of these individual estimates gives an overall corrected estimate of the parameter of interest and the 2.5^{th} and 97.5^{th} percentiles give a 95% simulation interval. However, this interval does not take into account the sampling error introduced by estimating the parameter of interest (regression coefficient or log odds ratio in this thesis). To take account of this, it is necessary to subtract from each corrected estimate the product of the observed standard error of the uncorrected estimate (obtained in a naïve analysis using the misclassified exposure) and a randomly selected value from the standard normal distribution [N(0,1)] (Fox et al. 2005, Greenland 2003). (Note that, since this product has a mean of zero, the same result would be achieved if the product was added to – rather than subtracted from – each estimate). The median of these corrected estimates and the 2.5^{th} and 97.5^{th} percentiles then give the overall corrected estimate and a 95% simulation interval that also takes account of the (sampling) error introduced by estimating the regression coefficient or log odds ratio.

An approach similar to this, called bootstrap imputation, was used by van Walravan (van Walraven 2017). In his study he generated (and validated) a (logistic) prediction model for the outcome of interest (a gold standard measure) using surrogate outcomes derived from administrative health records. The prediction model was used to generate probability estimates of the outcome for each individual in the study's validation subsample (the subset of individuals that was not used to generate the prediction model). Then, 1000 bootstrap samples of the study's validation subsample were generated and, for each individual within each of these bootstrapped datasets, a random number from a uniform distribution with parameters 0 and 1 was selected. If the random number was lower than the individual's predicted probability of the outcome then this person was classified as having the outcome. Similarly, if the random number was higher than their predicted probability, they were classified as not having the outcome.

2.2.2 Maximum likelihood methods

Estimates of sensitivity and specificity have also been incorporated into logistic regression models when the outcome (but not exposure) is misclassified (Magder and Hughes 1997). In this model, each individual is included twice, once as having the outcome of interest and once without, with weights determined by the probability that the individual has the outcome, given the value of their observed (misclassified) outcome and covariates. Odds ratio estimates are estimated using an expectation-maximisation (EM) algorithm. This method can be used when sensitivity and specificity are regarded as known, or when they are available from a validation study (Magder and Hughes 1997). Similarly, Edwards et al. used modified maximum likelihood estimators to correct for fixed values of sensitivity and specificity to take account of a misclassified outcome in Poisson regression models (Edwards et al. 2014). I do not use maximum likelihood methods in this thesis as I have a misclassified exposure (not outcome) variable.

2.2.3 Bayesian methods

Bayesian methods can also be applied in this context, either with (Chu et al. 2006, Richardson and Gilks 1993) or in the absence of a gold standard measure (Joseph et al. 1995, MacLehose et al. 2009, McInturff et al. 2004). Richardson and Gilks (Richardson and Gilks 1993) described how a Bayesian model for covariate measurement error can be subdivided into three sub-models:

1. The outcome (analysis) model – Y as a function of the true exposure X and other covariates C
2. The measurement (or misclassification) model – the misclassified exposure Z as a function of the true exposure
3. The exposure model – the true exposure X as a function of covariates C

This could also be approached as a missing data problem (Greenland 2009). In this case, the outcome model and the model for the missing true exposure are both specified as part of the Bayesian model (MacLehose et al. 2009). The latter model includes the observed (misclassified) exposure and can also include covariates but should not include the outcome. If misclassification depends on the outcome, this will feed into the misclassification model because both sets of parameters (those in the analysis model and those in the missing data (misclassification) model are jointly estimated as part of the same overall model.

If internal validation data are available, non-informative priors can be placed on all the parameters in the model. In contrast, in the absence of validation data it is necessary to specify informative priors on the prevalence of the true exposure, and the sensitivity and specificity (or predictive values) of the misclassified exposure (Corbin et al. 2017, Valle et al. 2015).

2.2.4 Multiple imputation

As mentioned in Section 2.2.3, when there is internal validation data, this can be treated as a missing data problem in which the (missing) true value of the exposure

(i.e. among individuals without the gold standard measure available) is imputed using multiple imputation (Cole et al. 2006, Edwards et al. 2013). Cole et al. compared multiple imputation to regression calibration – a method that can be used when a continuous covariate is subject to non-differential measurement error. They found that, in the case of a continuous covariate that was dichotomised for the analysis, MI performed as well as regression calibration and, in some scenarios, was more efficient (Cole et al. 2006). Similarly, Freedman et al. found that – for a continuous covariate subject to differential measurement error – MI and moment reconstruction (another method for correcting for measurement error in continuous covariates) were less biased than regression calibration but had larger variance (Freedman et al. 2008).

Multiple imputation has also been suggested as a way to simultaneously address missing data and measurement error (Blackwell et al. 2017).

It should be noted that MI is essentially an approximation to a full Bayesian analysis (Carpenter and Kenward 2013); as such, if the models used in MI are the same as those used in a Bayesian analysis, the results should be equivalent in large samples.

2.2.5 Comparison of methods to address misclassification

It should be noted that all the above methods assume that the misclassification probabilities are the same in the validation sample as they would be in the entire study once observed variables are taken into account – that is, the gold standard measure is MCAR or MAR conditional on the observed exposure, outcome and covariates (or other observed factors).

A few studies have compared different methods to address misclassification. Gilbert et al. compared fixed and probabilistic bias analysis, logistic regression and Bayesian methods to correct for outcome misclassification in a case-control study with external validation data. In simulations, estimates corrected using probabilistic bias analysis were less accurate that those corrected using logistic regression or fixed bias analysis (the Bayesian method was not used in the simulations). The four methods gave

similar estimates of the odds ratio when applied to a real dataset, although the Bayesian estimates were slightly closer to the null (and closer to the uncorrected estimate) than those obtained from the other methods (Gilbert et al. 2016).

Fixed bias analysis has been compared to bootstrap imputation (described above in Section 2.2.1 and similar in principle to probabilistic bias analysis) to correct prevalence and association estimates in the presence of misclassification, using internal validation data. Thus study found that using inaccurate bias parameters increased bias but, when perfectly accurate bias parameters were used, bias was eliminated; bootstrap imputation also reduced bias when the model used to predict the outcome was accurate (van Walraven 2017).

Using simulations, Corbin et al. compared fixed and probabilistic bias analysis (PBA) to a fully Bayesian method to correct for exposure misclassification in the absence of validation data. Fixed bias analysis performed poorly when the bias parameters used to correct the estimates differed from their true values. They concluded that a fully Bayesian analysis was the best method in terms of taking account of uncertainty in the parameters but – when there are no validation data – requires more assumptions in terms of prior information, since it is necessary to give a prior for the prevalence of the true exposure as well as for the misclassification parameters (Corbin et al. 2017).

More recently, Livingston et al. (Livingston et al. 2018) compared PBA, regression calibration and MI to correct for both non-differential and differential misclassification in a binary exposure. Regression calibration gave approximately unbiased odds ratios when the misclassification was non-differential and the sensitivity and specificity were both high (0.9). Regression calibration is not designed for differential misclassification and consequently performed poorly in this scenario. The authors found that neither MI nor PBA worked in small samples (n=200) but, in larger samples, MI produced estimates with little or no bias when misclassification was differential as well as when it was non-differential. PBA performed well when the sensitivity and specificity were high but gave biased estimates in the presence of

greater levels of misclassification (sensitivity and specificity = 0.6). Further, PBA resulted in wide simulation intervals.

Finally, Bartlett and Keogh compared Bayesian methods to regression calibration for a continuous covariate with replication data (rather than validation data)[1]. They also discussed the relative advantages of the Bayesian approach over MI (although did not directly compare MI to the Bayesian methods in their study). They argued that the Bayesian approach is preferable in the context of covariate measurement error because it can still be done in the absence of validation data whereas MI cannot. Further, they discussed the limitations of MI in this setting. Firstly, the imputation model for the missing true exposure may not necessarily be compatible with the analysis model. Secondly, they discussed the fact that Rubin's rules may not always perform well in this setting because the posterior distributions for the parameters in the analysis model are often skewed unless the sample size is large (Bartlett and Keogh 2016).

2.3 Literature review: use of linked data to address bias

To carry out my initial literature review I searched Web of Science (all databases) using the keywords shown in Box 2-2. This gave 465 references. There were twelve duplicates and I excluded 404 as not relevant after reading the titles. I searched the reference lists and citations of the remaining articles (and of the resulting relevant articles) and identified a further fifteen references. On further reading (of the 49+15=64 references), I found that only twenty examined bias in estimates of exposure-outcome associations; these are the papers I reviewed in full. Later I carried out a narrower search (also described in Box 2-2). This gave two more papers, and a search of their citations and reference lists gave two more.

[1] Replication data: repeat measurements on the same individuals using the same method of measurement; validation data: a subsample of individuals have a particular characteristic measured using both a gold standard method and an imperfect method of measurement.

Box 2-2: Keywords used in search strategy for use of linked data to address bias

Initial search

Topic = administrative data OR routine data OR administrative records
AND
Topic = data linkage OR record linkage OR linkage OR linked data NOT linkage disequilibrium NOT linkage analysis
AND
Topic = missing data OR non$response OR los? to follow$up OR attrition OR measurement error OR misclassification OR bias* OR imput*

Second (narrower) search

Topic = administrative data
AND
Title = bias or imput*

2.3.1 Use of linked data to address missing data

A number of studies have been carried out to evaluate bias due to non-response in which the outcome of interest comes from a linked data source (i.e. for all individuals in the study). Below I will summarise those studies that have used this information to examine the impact of non-response on estimates of exposure-outcome associations (Ferrie et al. 2009, Harald et al. 2007, Heilbrun et al. 1991, Knudsen et al. 2010, Lorant et al. 2007, Lundberg et al. 2005, Martikainen et al. 2007, Nilsen et al. 2009, Nohr et al. 2006, Nummela et al. 2011, Osler et al. 2008, Sogaard et al. 2004, Tin et al. 2014, Wigertz et al. 2010). Many of these studies also evaluated the impact of non-response on estimates of prevalence, incidence or mean levels of outcomes of interest; I will only summarise the findings relating to estimates of exposure-outcome associations.

In the Nordic countries the use of linked health and administrative data is facilitated by having national registries and unique identifiers that are consistent across all these datasets. This is reflected in the list of studies given in Table 2-1, where 9/13 (69%) of those identified were studies carried out in one of these countries. Most – but not all – of the studies found that estimates of the exposure-outcome relationship were very similar among respondents and non-respondents (Table 2-1). However, most of the

studies were looking at initial participation in a study as opposed to study drop-out / loss to follow-up and this will only generate bias in exposure-outcome estimates if the outcome (or an unmeasured cause of the outcome) is associated with participation. This is obviously possible in a cross-sectional study. In a longitudinal study, this is more likely when considering loss to follow up but could happen if there were an unmeasured risk factor for the outcome associated with study participation.

Table 2-1: Studies using linkage to provide outcomes on participants and non-participants

Authors	Country	Dataset(s) linked to	Response rate	Summary of findings in relation to exposure-outcome associations
Ferrie et al.	UK – England	NHS Central Register	76-86% **(follow-up)** of those who originally participated (initial participation was 73%)	The relationship between non-response and mortality did not differ according to job grade.
Harald et al.	Finland	Mortality registry	87% **(initial participation)**	The relationship between socio-economic position and mortality did not differ among participants and non-participants.
Heilbrun et al.	Hawaii (USA)	Cancer registry & hospital data	81% **(follow-up)** of those who originally participated (initial participation was 89%)	Various exposures were examined in relation to different cancers. Risk ratios among respondents and non-respondents were not significantly different, although numbers were small and confidence intervals quite wide.
Knudsen et al.	Norway	Disability pension registry	Although participation rate was 63% **(initial participation)**, only simulated ~5% missing exposure data	Simulated missing (MNAR) exposure data. Hazard ratios for participants were only slightly lower than those for all study participants (this comprised the 63% who responded initially, as only these individuals provided exposure data).
Lorant et al.	Belgium	Census data	61% **(initial participation)**	Found evidence for bias in estimates of socio-economic inequalities in self-rated health for some measures of socio-economic position but not others.
Lundberg et al.	Sweden	Various registries	53% **(initial participation)**	Odds ratios for the association between various socio-demographic variables and mental illness diagnoses were very similar among non-respondents and all invited individuals.
Martikainen et al.	Finland	Various city registries	67% **(initial participation)**	Relative rates of sickness absence by occupational social class were similar among respondents and non-respondents.

Table 2-1: Studies using linkage to provide outcomes on participants and non-participants

Nilsen et al.	Norway	Birth registry	44% **(initial participation)**	There was no evidence that estimates of eight exposure-outcome associations differed among participants and all those potentially eligible (whether or not actually invited).
Nohr et al.	Denmark	Birth registry	31% **(initial participation)**	Estimates of three exposure-outcome associations were very similar among participants and all eligible individuals.
Nummela et al.	Finland	Population registry	66% **(initial participation)**	Association between sex and poor health was reversed in respondents compared to non-respondents; association between income and poor health was only present among non-respondents. Other associations were similar in the two groups.
Osler et al.	Denmark	Psychiatric & prescription registries	66% **(follow up)**	There was no evidence that associations between early life characteristics and depression outcomes (diagnosis, prescriptions) differed among participants and all individuals invited for follow-up.
Sogaard et al.	Norway	Various registries	46% **(initial participation)**	Estimates of association between education and receipt of disability benefit were similar among respondents and all those invited to participate; association between country of birth and the same outcome was over-estimated among respondents.
Tin et al.	New Zealand	Insurance claims, hospital, mortality & police data	60% **(follow-up)** of the baseline group (initial participation 43%)	Adjusted hazard ratios for the association between a number of exposures and cycle crashes were similar among those who initially participated and those who responded to the follow up questionnaire.
Wigertz et al.	Sweden	Population registry	70% among controls; 79% cases **(initial participation)**	Odds ratio for the association between income and one type of brain tumour was different in participating vs non-participating individuals but similar for the other type of brain tumour studied.

Other studies have used linked data to calculate weights or carry out MI to address bias due to missing data. These results are summarised in Table 2-2. As above, I have not listed or summarised the results from studies that only examined bias in estimates of incidence, prevalence or mean values.

Faris et al. compared a "data enhancement" method to various imputation methods to investigate risk factors for mortality among cardiac patients (Faris et al. 2002). In the data enhancement method, information from equivalent variables from administrative (hospital discharge) data was used in combination with cohort data. So, for example, if an individual was recorded as having a particular risk factor in either dataset then it was coded as being present and if it was absent in both datasets it was coded as being absent. The imputation methods used included multiple imputation by chained equations. The odds ratios obtained after imputing using only the study data were similar after additionally including variables from the administrative data. However, they felt that their covariates were plausibly MAR (they did not use the administrative data to examine this).

Hebert et al. used blood pressure data from medical records to impute missing blood pressure measurements at follow-up in a clinical trial (Hebert et al. 2011). Although the focus of the paper was on bias in terms of the estimates of mean systolic and diastolic blood pressure, the authors did report regression coefficients for various predictors of blood pressure. The regression coefficients for some covariates were quite different when estimated using imputation models only using the study data compared to models that included blood pressure measures from linked medical records as auxiliary variables (for example, the coefficient for current smoking changed from -0.06 to 0.44).

Finally, Schomaker et al. used linked national mortality data, linked for a relatively large proportion of those taking part in HIV cohort studies in South Africa, to correct for missing outcome data (time to death or censoring) among those lost to follow-up (Schomaker et al. 2014). They used inverse probability weighting such that those subjects lost to follow-up for whom mortality data was available from the linked data

receive weights w>1. They also used MI and a combination of IPW and MI and compared the results, both on their dataset and through a simulation study. These methods all led to very different estimates of mortality compared to a complete case analysis or one that assumed non-informative censoring but estimates of exposure-outcome associations were only slightly altered (for example, the estimated hazard ratio for CD4 count category 25-50 compared to the reference group (<25) varied between 0.68 and 0.75).

The remaining studies listed in Table 2-2 are summarised in Section 2.3.2 since their primary focus was on measurement error.

Table 2-2: Studies using linkage data to carry out multiple imputation (MI) or weighting methods

Authors	Country	Dataset(s) linked to	Response rate	Summary of findings
Faris et al.	Canada	Hospital discharge data	Ranged from 73% to 97%	Imputed missing covariates (thought to be MAR). Results from imputations were similar to those obtained using "data enhancement"; they did not present a complete case analysis.
He et al.	USA	Linked between medical records and claims data	80% (medical records); 58% (claims)	This study is summarised in Table 2-3 because it was primarily focussed on addressing bias due to misclassification (under-reporting).
Hebert et al.	USA	Medical records	63% had complete follow up data	Used blood pressure (BP) measures abstracted from medical records to impute follow-up BP measurements in a trial. Also imputed based only on baseline BP. Some of the regression coefficients from the two imputation models were quite different. They did not give a complete case analysis. Used simulations to look at bias in mean BP but not regression coefficients and found that the addition of linked data reduced, but did not eliminate, bias in mean BP when BP data were MNAR.
Schomaker et al.	South Africa	Mortality register	3 cohorts: 73%, 78% and 90%	Used IPW and MI (plus a combination) to correct for missing survival data in HIV studies. Mortality estimates (Kaplan-Meier) varied quite substantially according to the method used and were under-estimated in the complete case analysis. Most exposure-outcome estimates were quite similar.
Yucel & Zaslavsky	USA	Linked between cancer registry and medical records	Unknown	This study is summarised in Table 2-3 because they were looking at this as a measurement error problem – misclassification due to under-reporting of cancer treatment in the registry.

2.3.2 Use of linked data to address misclassification

The literature search described in Section 2.3 covered the use of linked data to address both missing data and misclassification. As above, a relatively large number of studies have been carried in which a linked dataset was used to quantify the impact of measurement error on estimates of prevalence, incidence or the mean level of a response. Again, I have not included these studies in this review. Below I summarise only those studies that have used this information to examine the impact of measurement error on estimates of exposure-outcome associations. These studies are listed in Table 2-3. Of the eight studies listed, five linked between different administrative or routine datasets, rather than linking to an epidemiological study or survey.

In the studies by Yucel and Zaslavsky (Yucel and Zaslavsky 2005) and He, Landrum and Zaslavsky (He et al. 2014) the authors combined information from two data sources using Bayesian methods to simultaneously account for missing data and misclassification. In the former, they treated information from one source (medical records) – which was available for a subsample of individuals – as the gold standard. In the latter, they assumed both datasets were subject to misclassification and treated the true measure as a latent variable with 100% missing values, using Bayesian methods to impute (true) exposure status and modelling the under-reporting (sensitivity) of the two different measures of exposure with vague priors. In the first paper the main focus was on estimating the rate of exposure rather than comparing exposure-outcome estimates; in the latter they found that estimates of associations were substantially different in their Bayesian models compared to models in which the misclassified exposure status variables were used.

The remaining studies (Barry et al. 2013, Brochu et al. 2014, Kristensen and Irgens 2000, Macleod et al. 2002, Marshall et al. 2007, Randall et al. 2013) used an alternative data source to examine – but not correct for – measurement error, in the sense that they simply compared estimates obtained using one measure to estimates

obtained when using an alternative (linked) measure. All the studies except one found differences in exposure-outcome estimates.

Table 2-3: Studies using linked data to address measurement error

Authors	Country	Dataset(s) linked to	Summary of findings
Barry et al	Scotland	Scottish Morbidity Records (hospital data)	Used information from hospital records to compare risk reductions for cardiovascular endpoints comparing treatment groups in a clinical trial. Risk reductions were very similar when using hospital data to those in the trial.
Brochu et al	Canada	Government tax records	Linked census data (self-reported income) to tax records to estimate impacts of measurement error in estimates of (1) income and (2) inequality. Found that consent to linkage (thus missing information on true income) introduced bias in terms of measuring income inequality.
He et al	USA	Linked between medical records and claims data	Used Bayesian methods to impute true treatment status (assumed missing for all participants). Found substantially different estimates of associations compared to results using the misclassified treatment status variables.
Kristensen and Irgens	Norway	Linked between agricultural census and birth registry	Compared data on maternal recall of previous pregnancy loss to actual losses recorded in birth records. Found that estimates of predictors of loss were attenuated when using (misclassified) maternal recall compared to birth records, although numbers were small and confidence intervals wide.
Marshall et al	New Zealand	Linked between a GP database and the National Health Index database	Found differences in rate ratios for CHD admissions by ethnicity depending on which source of data was used for determining ethnicity.
Macleod et al	Scotland	Scottish Morbidity Records	Found associations between self-reported stress and symptoms of heart disease but no association with more objective measures (hospital admissions/mortality).
Randall et al	Australia	Linked between several administrative datasets.	Used various algorithms to combine information on Aboriginal status from various data sources to look at the impact of rate ratios for admissions and mortality. Different algorithms led to differences in the rate ratios.
Yucel & Zaslavsky	USA	Linked between cancer registry and medical records	Used medical records as the gold standard measure to impute true treatment status in the cancer registry (assumed to be subject to under-reporting). Focus was on estimating true treatment status rather than effect of treatment on outcome.

2.4 Participation in cohort studies

Galea and Tracy carried out a review of participation in epidemiologic studies (Galea and Tracy 2007). They were considering participation both in terms of initial participation as well as dropout. They reported that participation in epidemiologic studies has been shown to be associated with a number of socio-demographic factors, including sex, ethnicity, employment status, marital status, socio-economic position, and education level, with females, married individuals and those of higher SEP/education being more likely to participate. Since socio-economic position is – in general – associated with poorer health outcomes, it follows that participation will also tend to be lower among those with poorer health status. However, studies have shown that the presence of psychiatric disorders is associated with dropout in longitudinal studies even after taking account of socio-demographic factors (Bjertness et al. 2010, de Graaf et al. 2000). The results for ethnicity have been inconsistent, with some studies finding white individuals more likely to participate, but others finding higher participation among non-whites. In the review, it was reported that studies have also found that individuals who take part in risk behaviours such as smoking, drug and alcohol use are less likely to take part (Galea and Tracy 2007).

Participation in ALSPAC has been shown to relate to similar socio-demographic factors, with females, white individuals, those with higher educational attainment, and those from higher socio-economic backgrounds being more likely to participate (Boyd et al. 2012). More recently, two studies have looked at whether genetic factors are related to ALSPAC participation. One found that individuals with a higher polygenic risk score for schizophrenia were less likely participate (Martin et al. 2016); the other found associations between polygenic risk scores for BMI, education, ADHD, depression, schizophrenia, smoking and other personal characteristics such as agreeableness (Taylor et al. 2018). This suggests that these variables could be MNAR.

2.5 Background to the exemplars

The exemplars were chosen to be questions of epidemiological interest in which there was a potentially modifiable exposure and one or more plausible mechanisms for the association of interest – i.e. the association could potentially be causal. In addition, I used the following criteria:

- The outcome or exposure of interest was (likely to be) missing not at random.
- There was a proxy for the missing outcome or exposure available in a linked dataset.

Further, I chose the exemplars so that I had one missing continuous outcome (IQ), one missing binary outcome (depression) and one missing exposure (teenage smoking). This ensured that I was covering a range of different scenarios, particularly in terms of missing data.

2.5.1 Breastfeeding and IQ

The UK has one of the lowest rates of breastfeeding in the world, with only 34% of babies receiving any breast milk at 6 months and less than 1% at 12 months (compared to over 90% at 12 months in a number of low and middle income countries (LMIC)) (Victora et al. 2016).

IQ is a strong predictor of educational attainment (Deary et al. 2007) and higher educational attainment has been shown to have a positive impact on health outcomes across the lifecourse (Marmot 2010). Understanding whether there is a causal association between breastfeeding and IQ is thus important because this adds to the evidence regarding the its beneficial effects, thereby strengthening the need for interventions to increase breastfeeding rates.

Robust evidence suggests that breastfeeding is associated with higher childhood and adolescent cognition (Anderson et al. 1999, Horta et al. 2015, Horta 2013, Kramer et al. 2008, Lucas et al. 1992). In a recent meta-analysis of 18 studies (Horta et al. 2015),

the pooled estimate of the mean difference in IQ for those breastfed compared to those who were not was 3.44 points (95% CI: 2.30, 4.58).

It has been argued that the association could be due to confounding by parental cognitive outcomes, socio-economic factors, or differences in parenting behaviour and interactions between the mother and child (Gibbs and Forste 2014, Horta et al. 2015, Jacobson et al. 2014, Walfisch et al. 2013). Indeed, in the meta-analysis by Horta et al., the estimated mean difference in IQ (comparing breastfed children to those not breastfed) among studies that adjusted for maternal IQ was lower than the combined estimate from all studies: mean difference = 2.62 (1.25, 3.98) compared to 3.44 in all studies (Horta et al. 2015). However, the association has been observed in two randomised controlled trials (Kramer et al. 2008), as well as in a study which compared results from ALSPAC to those from a cohort study in Pelotas, Brazil (Brion et al. 2011). The rationale for comparing estimates from a high-income country to those from a LMIC country is that the confounding structures are likely to differ (and did in this case), implying that the observed association is less likely to be due to residual confounding. In addition, studies have shown that breastfeeding is positively associated with white matter development (Deoni et al. 2013, Isaacs et al. 2010); it is hypothesised that this could due to the presence of long-chain fatty acids in breast milk (Deoni et al. 2013).

2.5.2 Smoking in pregnancy and offspring depression

Some evidence suggests a substantial increase in rates of depression and anxiety among children and adolescents in the UK in the past few decades (Collishaw et al. 2010) with a recent study finding that 24% of girls and 9% of boys report high levels of depressive symptoms at age 14 years (Patalay and Fitzsimons 2017). Depression has long term consequences, impacting negatively on education, employment, quality of life and both physical and mental health (Hankin 2006). Indeed, depression has been shown to be one of the leading causes of disability and premature death worldwide (Ferrari et al. 2013). Given this, it is vital to establish the causes of depression in order to be able to design and target appropriate interventions.

That smoking in pregnancy is harmful – to the unborn baby as well as its mother – is not in question (Smoking in Pregnancy Challenge Group 2018). In 2010 the UK government set a target to reduce smoking in pregnancy rates to 11% by 2015. This target was met, but the rate now remains relatively stable (10.8% in the year to April 2018) and the new target (6% by 2022) may not be reached (Smoking in Pregnancy Challenge Group 2018).

Some studies have found an association between exposure to maternal smoking during pregnancy and offspring internalising behaviour problems (Ashford et al. 2008, Indredavik et al. 2007, Moylan et al. 2015), including depression (Ekblad et al. 2010, Menezes et al. 2013); others have found no association (Brion et al. 2010, Dolan et al. 2016, Lavigne et al. 2011, Monshouwer et al. 2011, Orlebeke et al. 1999). Plausible mechanisms for such an association have been suggested. For example, nicotine is known to affect neurological development and, in particular, has been shown to impact on serotonin, dopamine, and noradrenaline transmission systems (Moylan et al. 2013), which are thought to play an important role in depression (Ressler and Nemeroff 2000).

In a recent cross-cohort study that included ALSPAC (Taylor et al. 2017), an overall association with offspring depression was found: OR = 1.20, 95% CI (1.08, 1.34) (Figure 2-1). This was a combined estimate across four cohort studies. However, within this same study the authors used data from a fifth cohort which included 258 pairs of siblings that were discordant for maternal smoking; in this analysis there was no evidence for an association: OR=1.03 (0.77, 1.36) (Taylor et al. 2017). The authors of this study concluded that the observed association between maternal smoking in pregnancy and offspring depression is likely to be due to unmeasured confounding by socio-economic position and other parental characteristics (Taylor et al. 2017).

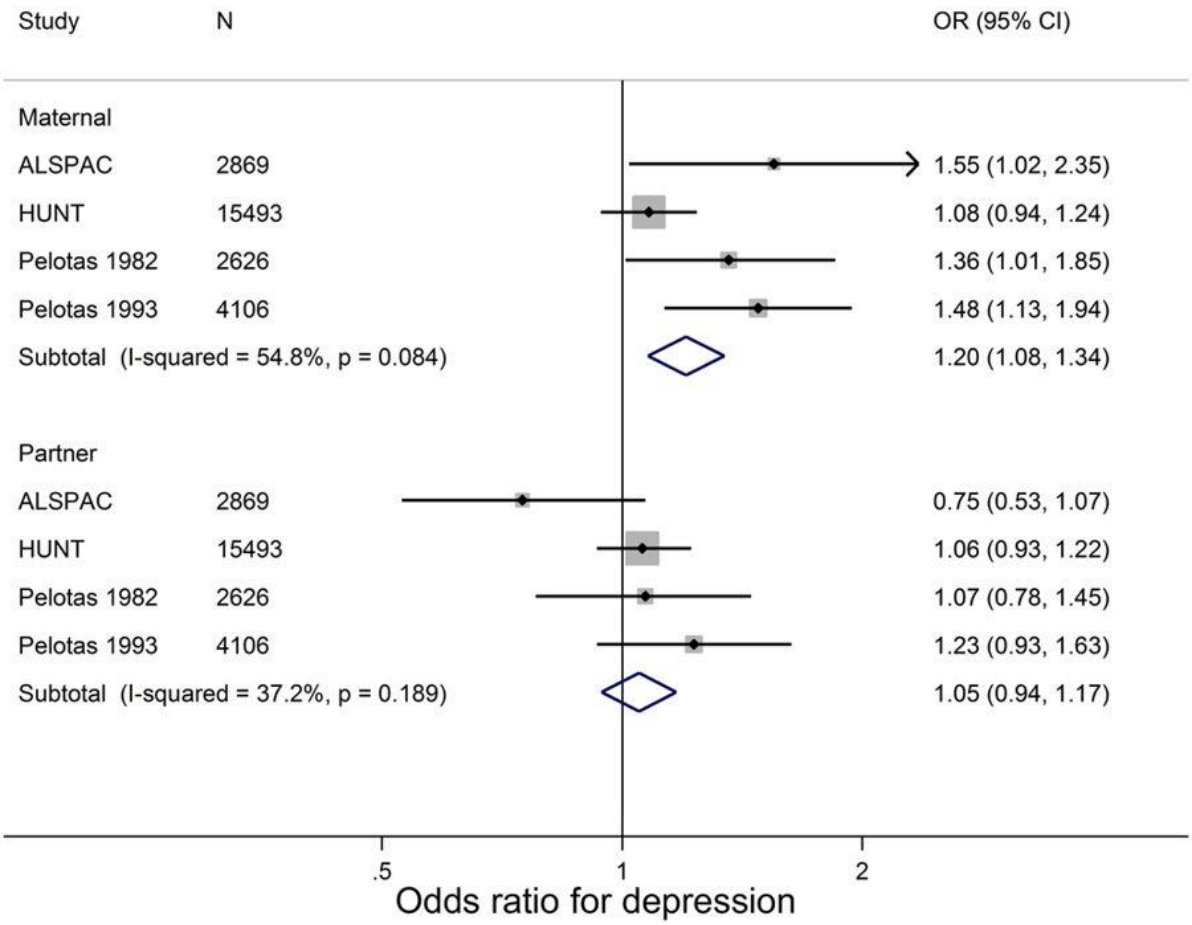


Figure 2-1: Combined estimates of the effect of maternal and paternal smoking during pregnancy on offspring depression

FROM: Maternal smoking in pregnancy and offspring depression: a cross cohort and negative control study (Taylor et al. 2017) (https://www.nature.com/articles/s41598-017-11836-3). Published open access under a CC BY license (Creative Commons Attribution 4.0 International License).

2.5.3 Teenage smoking and educational attainment

As stated above, educational attainment is an important predictor of many health outcomes, with higher attainment leading to better health and longer life expectancy (Marmot 2010). Teenage smoking in England has been declining since the 1980s: the percentage of 11 to 15 year olds who classify themselves as current smokers has dropped from around 20% in the early 1980s to an estimated 6% in 2016, although remained steady between 2014 and 2016 (NHS Digital 2018). Regardless of any potential effect on educational attainment, the importance of reducing rates of

smoking among teenagers has already been recognised. However, if teenage smoking is causally linked to poorer educational attainment, this would add further weight to the argument.

In observational studies, teenage smoking has been previously shown to be associated with dropping out of school early (Bray et al. 2000, Lynskey et al. 2003) and on educational attainment (Busch et al. 2017, Koivusilta et al. 2003, Pennanen et al. 2011, Stiby et al. 2015). It has been hypothesised that this association could be mediated through effects on prefrontal cortical function (Galvan et al. 2011) or through an impact on psychosocial problems (Busch et al. 2017). However, as with effects of prenatal exposure to smoking, it has also been suggested that the association may be due to confounding, particularly by socio-economic position (Koivusilta et al. 2013) as well as familial, peer and school factors (Glendinning et al. 1995) or that smoking and other risk behaviours may simply signify a negative attitude towards education (Busch et al. 2017). It has also been suggested that the association may be in the other direction, with lower educational attainment resulting in an increased risk of smoking initiation and lifetime use (Gilman et al. 2003). Although it has been argued that this association could also be due to confounding (Farrell and Fuchs 1982, Gilman et al. 2008), evidence for causality has been found in a recent Mendelian randomisation study (Gage et al. 2018). Note, however, that the finding in this study could indicate an association in either direction because the genetic markers for educational attainment are a marker of lifecourse "exposure".

Chapter 3 Data sources

The data used in this research come from a birth cohort study, the Avon Longitudinal Study of Parents and Children (ALSPAC), which is described in detail in a cohort profile paper (Boyd et al. 2012). A brief summary of ALSPAC, details of data linkage relevant to this thesis, and information about the variables used in the analyses are given below.

3.1 Summary

ALSPAC is a prospective observational study which recruited pregnant women in 1990-1992. All pregnant women living in one of three health districts within the former county of Avon with due dates between 1st April 1991 and 31st December 1992 were eligible to take part. The catchment area is shown in Figure 3-1 and comprised the three health districts that became the Bristol and District Health Authority: Southmead, Frenchay and Bristol and Weston District Health Authorities.

Parts of this chapter have been previously published: Cornish RP et al. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. Int J Epidemiol 2015; 44(3):937-45 https://doi.org/10.1093/ije.dyv035; Cornish RP et al. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. Emerg Themes Epidemiol 2017 14:14. https://doi.org/10.1186/s12982-017-0068-0; and Cornish RP et al. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. BMJ Open 2016 6:e013167. https://doi.org/10.1136/bmjopen-2016-013167. These papers were all published open access under a CC BY license (Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0/). Changes have been made to some of the published text.

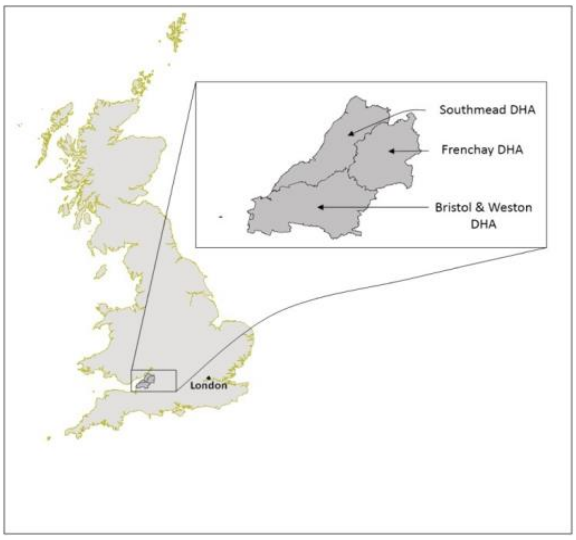


Figure 3-1: The ALSPAC catchment area

Among the eligible pregnancies (n=20,248), 14,541 were recruited in 1990-1992; these resulted in 14,062 live-born children. A further 713 children were recruited in later years. Thus, the total number of enrolled children was 14,775. Of these, 14,762 were singletons or twins, 14,683 of whom were alive at one year. The enrolled participants have been followed up regularly since birth, with data collected through various mechanisms:

- Self-completed questionnaires for the mothers (about themselves and their index child(ren)), their partners and the index children.
- Study clinics attended at the University of Bristol, in which various physical, cognitive, psychological and other measurements were recorded.
- Linkage to external datasets (see Section 3.2).

The sample included in the analyses presented in this thesis vary according to the exemplar question. The participation analyses and the analysis for Exemplar 1 (breastfeeding – IQ) included only core cases (those enrolled in the original recruitment phase in 1990-1992). The former analysis only included core cases because my intention was to carry out a complete case analysis and only core cases

would have information on baseline covariates; further, subsequent recruitment did not start until after 7 years, so there would be missing information on participation up to this age. For the latter analysis, I could potentially have included non-core cases but breastfeeding information (as well as all baseline covariates apart from sex and mother's age) was completely missing for these individuals.

The other analyses (Exemplar 2: smoking in pregnancy – offspring depression and Exemplar 3: teenage smoking – educational attainment) also included those who enrolled subsequently. I included non-core cases in Exemplar 2 because, although smoking in pregnancy was only measured on core cases (so no non-core cases would be in the complete case analysis), a greater proportion of non-core cases than core cases (38% compared to 31%) had ALSPAC data on depression at 18. All analyses exclude triplets and quadruplets, individuals who died before age one year as well as any individuals who had withdrawn from the study. Since individuals can withdraw at any time and I received the datasets used in this thesis at different times, the numbers included in the different exemplars vary slightly. This is summarised in Table 3-1.

Table 3-1: Numbers included in each analysis in this thesis

Analysis	Inclusion criteria[1]	Date dataset received	Overall N (males; females)
Participation in ALSPAC	Enrolled at baseline only	October 2016	13,972 (7,217; 6,755)
Inclusion in GP data	Enrolled at baseline; had valid NHS number	February 2017	13,864 (7,159; 6,705)
Exemplar 1: breastfeeding and IQ	Enrolled at baseline only	April 2014	13,975 (7,219; 6,756)
Exemplar 2: smoking in pregnancy and offspring depression	Had valid NHS number	February 2017	14,566 (7,474; 7,092)
Exemplar 3: teenage smoking and educational attainment	Had valid NHS number	June 2017	14,566 (7,474; 7,092)

1. All analyses included only singletons and twins who were alive at one year and who had not withdrawn from the study (i.e. requested that their data no longer be used) by the date specified in the table.

3.1.1 ALSPAC eligible versus enrolled individuals

Because there was no sampling frame from which to identify eligible pregnancies, ALSPAC recruitment was opportunistic via maternity services. However, the eligible population was subsequently identified through maternity, birth and health records {Boyd, 2012 #6784}. This (eligible) population has been compared to the enrolled cohort with respect to a limited number of characteristics. Enrolled mothers are less likely to be younger (<25) and less likely to have a partner with a lower occupational social class {Boyd, #8380} [unpublished work].

3.2 Data linkage in ALSPAC

When the pregnant women were enrolled into ALSPAC they were informed that the study planned to abstract data from their medical records; they were provided with an opportunity to opt out of this linkage (Fraser et al. 2013). Using these permissions, antenatal data and some data from birth records were abstracted. In addition, in 2008 the NHS Information Centre (NHS IC, subsequently called the HSCIC, Health and Social Care Information Centre and now NHS Digital) linked ALSPAC participants with the NHS Central Register, with a 99% match rate; this was done on the basis of NHS ID number, name, date of birth, and postcode using deterministic linkage (Boyd et al. 2012).

In 2009 the Project to Enhance ALSPAC through Record Linkage (PEARL) was set up to develop and establish mechanisms for linking to a range of different health and administrative datasets. Since ALSPAC needed to seek consent from the index children for continued enrolment (hereafter referred to as re-enrolment) in the study once they had reached legal adulthood (age 18 years), it was decided to combine this consent campaign with the request (via PEARL) for ALSPAC to link to and use participants' routine records; this was done via a postal campaign. All children from ALSPAC-enrolled families, except those who had died, withdrawn from the study after the age of 14 years, could not be traced, or were flagged on the ALSPAC administrative database as being not contactable due to a range of family or health circumstances (for example, known to lack the capacity to consent) were sent an

information pack about the study and details of the proposed linkages. The information pack included a covering letter, a short summary leaflet, a 32-page booklet and a consent form. The consent request was structured on an 'opt-out' basis, although ALSPAC requested that their preference was for an explicit consent decision. An audio version (CD) of the materials was also included in the pack and, in addition, the complete pack was made available on the ALSPAC website. Participants were given the opportunity to ask questions via telephone or email and were asked to return the consent form using a pre-paid envelope. The consent form requested consent for (a) study re-enrolment and (b) linkage to health records, education records, criminal convictions and cautions data, and benefits and earnings data. Consent was asked for each of these separately.

At the time of writing, there were a total of 13,652 enrolled singletons or twins who were alive at one year and who had been sent a consent pack (or received one in person at an ALSPAC event or study clinic). Of these, 619 packs were returned to ALSPAC with "addressee unknown"; thus, 13,033 individuals were assumed to have received fair processing materials. Among these, 12,595 (96.6%) individuals had either not responded or had explicitly consented to linkage to their health records. Similarly, among the 13,033 who received fair processing materials, 12,670 (97.2%) had either not responded or had explicitly consented to linkage to their education records.

Current EU legislation requires studies to collect explicit consent in order to access 'sensitive' identifiable records, including all health records. However, ALSPAC also gained permission (under the provisions of Section 251 of the NHS Act) from the Health Research Authority's Confidentiality Advisory Group (HRA CAG) to access medical records of individuals enrolled in the study who did not respond to this consent campaign. One condition of this approval was that certain data deemed to be particularly sensitive, including all mental health data, would be excluded from any data extracts unless subsequently approved by the committee on a study-specific basis. Thus, for the exemplars used in this thesis involving mental health data, I

applied for and gained approval from HRA CAG to extract data on all those who did not respond to ALSPAC's consent request (in addition to those providing explicit consent).

The sections below give more details about the linkages relevant to this thesis.

3.2.1 Linkage to the National Pupil Database (NPD)

The National Pupil Database is a longitudinal database containing attainment and other school and pupil-level data for children attending schools in England (https://www.gov.uk/government/collections/national-pupil-database). Linkage between ALSPAC and the NPD was originally conducted in 2002 by The Fischer Trust – an independent charity – in the role of a trusted third party. The linkage was deterministic and was carried out using name, date of birth, sex and address. Contribution of data to the NPD is compulsory for schools that follow the national curriculum. Independent (fee-paying) schools are not obliged to follow the national curriculum and therefore any individual who attended an independent school at the time of the linkage in 2002 would not have been linked unless their school contributed data voluntarily. Conversely, attainment data at General Certificate of Secondary Education (GCSE) level (not earlier) for those who were in a state maintained school at the time of the linkage but subsequently attended an independent school will be present as long as they sat GCSEs. Altogether, 14,007 enrolled singleton and twins alive at one year were originally linked to at least one dataset within the NPD. However, the linkage consent campaign was carried out after this linkage had taken place; out of these linked individuals, 12,670 had been sent fair processing materials and had not explicitly dissented to linkage to their education data.

3.2.2 Linkage to GP data

The NHS Wales Information Service (NWIS) and the Health Informatics Research Unit at the University of Swansea have established a method through which individual level data from multiple sources can be linked and analysed in a secure setting; this

includes data from primary care electronic patient records. This method was developed as part of the Secure Anonymised Information Linkage (SAIL) project, which has been described in detail before (Ford et al. 2009). ALSPAC, working in conjunction with the SAIL team, developed two methods to extract GP records which took advantage of the existing SAIL infrastructure. These are described below.

3.2.2.1 2012: pilot extraction (consenters only)

NHS Digital maintains a linkage between ALSPAC and the NHS Patient Demographic System as part of the ALSPAC 'Flagging and Tracing' study. Through this linkage, ALSPAC was provided with the current GP practice details for consenting individuals. ALSPAC then contacted the GPs seeking assent for the extraction of participants' records. From these, they identified practices that used a software system supplied by Egton Medical Information Systems Ltd (EMIS) (http://www.emis-online.com) or had installed practice record reporting software developed by Apollo Medical Systems Ltd (Apollo) (http://www.apollo-medical.com/index.php). ALSPAC commissioned EMIS and Apollo to extract the coded values of the participants' records (free text components were not extracted) from these practices.

This extraction was based on the 2,806 children who had provided explicit consent to linkage to their health records by October 2012 and were linked to one of 523 GP practices. By August 2013, ALSPAC had gained the authorisation to extract records from 290 (55%) of these practices (16 (3%) had refused authorisation by this date and contact was ongoing with the remaining 217 (42%)). Among these 290 practices, 264 used either EMIS or Apollo software, or both. ALSPAC extracted the records of 2,249 participants from 181 practices (extracts from the remaining 83 practices could not be conducted due to technical/governance issues relating to the Apollo extract system or the underlying practice software system).

3.2.2.2 2016: main extraction

The NHS South West Commissioning Support Unit (SWCSU) has developed a governance framework and data extraction mechanism which secured opt-in assent

from GP practices for the extraction of records and their use for SWCSU-approved purposes (for example, to support Connecting Care, a regional summary care record). Invitations to participate in this system were made to all practices in the Bristol, North Somerset, Somerset and South Gloucestershire (BNSSSG) clinical commissioning group. (Note that the area covered by the BNSSSG is roughly equivalent to what was formerly the county of Avon, the ALSPAC recruitment area.) The extraction mechanism is provided by EMIS, which supplies software systems to the majority of practices in the BNSSSG area. ALSPAC gained approval from the SWCSU Security and Informatics Group to extract participants' GP records. SWCSU informed all participating practices about this agreement and gave them the opportunity to opt out of this data sharing.

For both the pilot study and the main extraction, the methods after extraction were identical. The extracted records were anonymised and securely transferred into the SAIL infrastructure by the NWIS, following NHS encryption and security standards. The anonymisation took place using SAIL's "split file" method: once extracted, the GP software system supplier split the data into (1) a file containing identifiers and (2) a clinical file. They then assigned corresponding records within these files the same unique but otherwise meaningless batch number. The file of identifiers was encrypted and sent over the NHS N3 secure network to NWIS. NWIS was also sent ALSPAC unique ID and demographic matching fields on individuals (including NHS ID) with appropriate permissions (consent or Section 251 support). The two files were matched using an automated process that converted all the identifiers into an anonymised linking field (ALF) unique to the individual. NWIS then produced an output file containing the batch number, ALF and ALSPAC ID and excluding any externally meaningful identifiers such as NHS ID. This linking file was kept by NWIS and was used to add ALFs to the clinical files before transferring them into ALSPAC's folder within the SAIL infrastructure. In a similar way, the ALSPAC data linkage team sent ALSPAC data files (also containing ALSPAC ID) to NWIS. Using the linking file, NWIS replaced ALSPAC ID with ALF and transferred the resulting dataset into the ALSPAC folder within the SAIL infrastructure. The data were stored and analysed in a

study-specific folder with restricted access; no data could be removed from the SAIL infrastructure.

Of the 14,683 enrolled singletons and twins alive at one year, 13,033 (89%) were sent fair processing materials and thus given the opportunity to either consent or dissent to health data linkage. Among these, 438/13,033 (3.4%) dissented to linkage to their health records. ALSPAC had no record of NHS ID for 23 of the remaining 12,595, leaving 12,572 where linkage to GP records was attempted. GP records were extracted for 11,678 (93% of individuals where linkage was possible; 80% of enrolled singletons and twins alive at one year).

3.3 ALSPAC study data used in this thesis

The study data used in this thesis come from both questionnaires and clinics. (Note that I describe all data that came from linkage to external datasets separately, in Section 3.4.) The timing of the clinics and questionnaires in which exposure and outcome data were collected, together with the variables obtained from each, are shown in Table 3-2; details about covariate data are described in Section 3.3.3 and summarised in Table 3-3.

Table 3-2: Exposure and outcome variables for the three exemplars

	Age / time point					
	P	<18 mths	12-14y	15y	16y	18y
Maternal smoking in pregnancy	M					
Breastfeeding		M				
Teenage smoking			Cl, Cc			
IQ (WASI)				Cl		
Depression and anxiety (CIS-R)						Cl
Educational attainment					L	

Key:
P: during pregnancy
M: carer-based questionnaire (usually completed by the mother)
Cc: child completed questionnaire
Cl: variable measured in ALSPAC study clinic
L: variable obtained via linkage (including manual abstraction of records) to health or administrative datasets; these variables described in Section 3.4.

3.3.1 Outcome variables

3.3.1.1 Depression

Depression was measured using a self-administered, computerised version of the revised Clinical Interview Schedule (CIS-R) (Lewis et al. 1992) completed during a study clinic attended when the children were aged between 17 and 18 years. The CIS-R asks questions about fourteen symptoms of depression and anxiety (somatic symptoms, fatigue, sleep problems, irritability, worries about physical health, depression, depressive ideas, worries, anxiety, phobias, panic, compulsive behaviours, obsessive thoughts, and forgetfulness or concentration problems), and can be used to generate a total score as well as to assign ICD–10 (International Classification of Diseases, 10th revision) diagnoses of depression and anxiety disorders, including mixed anxiety and depression (Bebbington et al. 2007, Brugha et al. 1999, Lewis and Araya 2001). In ALSPAC, questions about obsessive and compulsive behaviours were omitted.

The outcome used in this thesis was whether or not an individual met the criteria for a diagnosis of depression.

3.3.1.2 IQ at age 15 years

IQ was measured at a study clinic attended when the children were 15 years old using the Wechsler Abbreviated Scale of Intelligence (WASI) (Wechsler 1999). For practical reasons, only two of the four WASI subtests (the vocabulary and matrix reasoning subtests) were administered; together, these provide a measure of general cognitive functioning.

3.3.2 Exposures

3.3.2.1 Breastfeeding

The breastfeeding information used in this thesis was collected via questionnaires filled in by the mothers at 4 weeks, 6 months and 15 months. The exact questions

used at each time point are given in Appendix A, Section 1. Duration of breastfeeding was categorised, with those who reported breastfeeding for less than one month combined with those who never breastfed.

3.3.2.2 Smoking in pregnancy

Mothers were asked about smoking habits when they were 18 weeks and 32 weeks pregnant and when their baby was around 8 weeks old. Again, the questions used are given in Appendix A, Section 1. At 18 weeks they were asked whether they had smoked regularly in the first three months of pregnancy and whether they had smoked regularly in the past two weeks. They were also asked how much they smoked per day during these periods. At 32 weeks they were asked how much they smoked per day. Finally, when the baby was 8 weeks old they were asked how much they had smoked per day during the last two months of their pregnancy. An individual who reported smoking at any of these time points was classified as having smoked during pregnancy; anyone who reported not smoking in all these questionnaires was classified as not having smoked during pregnancy; otherwise smoking status was classified as missing.

3.3.2.3 Teenage smoking

The smoking variables used in this analysis come from self-reported data collected during study clinics at ages 12, 13 and 15 years and a questionnaire administered at 14 years and, for a subsample, serum samples (in which cotinine levels were measured) taken at the study clinic attended at 15 years.

Three smoking variables were used in the analysis:

(1) Whether an individual had ever smoked by age 15 (yes/no)
    Individuals were classified has having ever smoked if they reported ever smoking at either 12, 13, 14, or 15 years; never having smoked if they reported never smoking at all these time points; and missing otherwise.
(2) Frequency of smoking (never, <daily, daily) at age 15
(3) Cotinine (>9.5 ng/ml vs ≤9.5 ng/ml)

This cut-off has been used previously in ALSPAC (Stiby et al. 2015) and has been shown to fall within the range of cut-off values with high sensitivity and specificity for detecting smokers (Jarvis et al. 2008).

3.3.3 Covariates

3.3.3.1 Baseline covariates

A number of socio-demographic measures thought to be potential confounders and/or predictive of non-response were collected during pregnancy and at birth. These are listed in Table 3-3.

Table 3-3: Baseline covariates used in this thesis

Factor	Timing	Further information / categories
Maternal factors		
Maternal education	32 weeks	O level/lower, A level, degree/higher
Ethnicity	32 weeks	White/non-white
Maternal age	At birth	<20, 20-24, 25-29, 30-34, 35+ years
Parity	18 weeks	0, 1, 2+
Age at first birth	18 weeks	<20, 20-24, 25-29, 30+ years
Depression score	18 weeks	Edinburgh Postnatal Depression Score (EPDS) (0-30)
Anxiety score	18 weeks	Crown Crisp Experiential Index (CCEI) (0-16)
Ever smoked	18 weeks	Yes vs no
Marital status	8-42 weeks	Married vs single/widowed/divorced/ separated
Alcohol use in early pregnancy	18 weeks	No if reported having <1 glass of alcohol per week; yes otherwise
Drug use in early pregnancy	18 weeks	Yes vs no (any)
Paternal factors		
Paternal education	32 weeks	Categories as per maternal education
Ever smoked	18 weeks	As for maternal measure
Depression score	18 weeks	As for maternal measure
Family / housing-related / combined parental factors		
Occupational social class	32 weeks	The higher of maternal and paternal social class; categorised as I-IIINM and IIIM-V[1]
Housing tenure	8-42 weeks	Owned/mortgaged, private rented, other
Number of rooms in house	8-42 weeks	Exludes bathrooms/toilets
Crowding index	8-42 weeks	Number of people/number of rooms: ≤0.5, 0.51-0.75, 0.751-1, >1
Telephone in house	8-42 weeks	Yes, no/incoming only
Use of car	8-42 weeks	By mother/carer or partner
Double glazing	32 weeks	Full/partial, none
Financial difficulties	8-42 weeks	Score (0-15), with higher score indicating greater financial difficulties
Adversity index[2]	8-42 weeks	Score (0-11), with higher score indicating greater adversity
Child factors		
Sex	At birth	

1. I-IIIN: professional, managerial, and non-manual skilled occupations; IIIM-IV: manual skilled occupations, semi-skilled and unskilled occupations.
2. Adversity index: a composite measure of social adversity taking into account a variety of factors, including items relating to housing, financial difficulties, family size and problems, maternal age and education, availability of social and financial support, substance abuse, and crime.

3.3.3.2 Additional covariates included in specific exemplars

In addition to the baseline covariates, an additional variable – not measured at baseline – was included in Exemplar 2 (smoking in pregnancy and offspring depression) as an extra auxiliary variable: the emotional difficulties score from the

Strength and Difficulties Questionnaire (SDQ), completed by the mother when the child was 7 years old.

3.4 Variables from linked datasets

3.4.1 Proxy variables obtained via linkage to the NPD

Several attainment variables were used from the NPD – from the Key Stage 2, Key Stage 3 and Key Stage 4 (KS2, KS3 and KS4, respectively) attainment data. A Key Stage is a period of the education system in the UK (except Scotland). KS2 covers ages 7-11 (Years 3 to 6), KS3 ages 11-14 (Years 7-9), and KS4 refers to the two school years attended when children are aged 14-16 years (Years 10-11). At the end of each of these Key Stages children sit statutory assessment tests. At KS2 and KS3 these are National Curriculum tests in English, maths and science (science tests were discontinued at Key Stage 2 in 2010 but were sat by ALSPAC children); in KS4 pupils typically take GCSE or equivalent vocational courses and are graded for each of these. The variables used in this thesis were:

KS2 and KS3 attainment scores

These variables were derived from the National Curriculum test results. In English, pupils have a reading test which is scored out of 50 and a writing assessment (out of 50). In maths there are two papers, both scored out of 40 in KS2 and out of 60 in KS3, and a mental arithmetic test (out of 20 in KS2 and 30 in KS3); finally, in science there are two papers, both scored out of 40 in KS2 and out of 90 in KS3. The attainment scores were obtained by adding together the English, maths and science scores, thus giving a maximum possible attainment score of 280 in KS2 and 430 in KS3.

KS4 capped point score

Each GCSE grade is equivalent to a specified number of points, with higher scores indicating higher attainment. The capped point score is calculated as the total score of an individual's top eight qualifications ranked in terms of points. The number of GCSE qualifications taken by a pupil can vary but the majority of pupils take at least

eight courses; the capped score therefore provides a more standardised measure of attainment than an individual's total score.

KS4: number of A*-C grades

This was dichotomised as <5 or 5 or more A*- C grades.

Two further variables from the NPD were also used:

Percent absence in Year 11 (authorised and unauthorised absences)

This was calculated on the basis of all schools attended during Year 11 (where >1 school was attended).

SEN (Special Educational Needs) status in Year 11

This was categorised as: no SEN, school action support, or statement of SEN. SEN Statements – now replaced by Education, Health and Care Plans (EHC Plans) are a description of a child's special educational needs and any additional support that they should receive in school. A child whose needs cannot be met by their school may receive a statement; this is determined after the child has undergone an assessment by the local authority (https://www.nhs.uk/Livewell/Childrenwithalearningdisability/Pages/Education.aspx)

3.4.2 Proxy variables obtained via linkage to GP data

Because the data in routine health datasets consist of one or more codes relating to diagnoses, symptoms, tests, and so on, associated with a date on which, for example, the person consulted the GP or attended hospital, variables need to be derived by means of an algorithm. These are generally comprised of one or more sets of codes as well as a time frame. For example, if the aim was to measure the prevalence of asthma at age 8 years, one algorithm might state that an individual would need to have at least one record of an asthma diagnosis while aged 8 to be included in the numerator; an alternative definition – or algorithm – could additionally include individuals who had a historical record of a diagnosis (i.e. before age 8) and received at least one prescription for asthma medication while aged 8. There are often several

– or many – potentially valid ways of defining an outcome or exposure in this way. In the following sections I will summarise algorithms used in previous research to define the exposures and outcomes relevant to my exemplars. If I modified these algorithms for my research questions, this is detailed in later chapters.

3.4.2.1 Defining depression using GP data

The extracted primary care data consisted of Read codes version 2 (5 byte) (https://digital.nhs.uk/article/1104/Read-Codes), together with associated dates. In an earlier study among adults, John et al. (John et al. 2016) identified sets of codes indicating diagnoses, symptoms and treatment (antidepressants, anxiolytics and hypnotics) for common mental disorders (CMD). For depression, this included the Read codes listed in Tables 1 to 3 in Appendix A; these codes were used to generate a number of different definitions of CMD involving diagnosis, treatment and symptoms. These were compared to a survey-based measure of CMD. Treatment, symptoms and diagnoses were classified as current if the code was recorded within six months either side of the survey date and historical if recorded more than six months prior to the survey date (John et al. 2016).

3.4.2.2 Defining smoking status using GP data

A number of studies have defined smoking status using Read codes. Two recent ones used a similar set of codes (Atkinson et al. 2017, Mukherjee et al. 2014). The codes used by Atkinson et al. are given in Tables 4 to 6 in Appendix A. Mukherjee et al. specified some additional Read codes (Mukherjee et al. 2014). Some of these either indicated passive smoking, drug smoking or tobacco chewing, or were specific to asthma or chronic obstructive pulmonary disease patients but the following codes did not fall in any of these categories: 137k. (Refusal to give smoking status), 137n. (Total time smoked), 137o. (Waterpipe tobacco consumption), 13p1. (Smoking status at 4 weeks), 13p2. (Smoking status between 4 and 52 weeks), 13p3. (Smoking status at 52 weeks), 13p50 (Practice based smoking cessation programme start date), 13p6. (Carbon monoxide reading at 4 weeks), 13p7. (Smoking status at 12 weeks), 8HBP.

(Smoking cessation 12 week follow up), 8IEo. (Referral to smoking cessation service declined), 8T08. (Referral to smoking cessation service), 9Ndf. (Consent given for follow-up by smoking cessation team), H3101 (Smoker's cough).

In Chapters 6 and 7 I discuss how I modified the above code lists for depression and smoking (respectively) for use in the work presented in this thesis.

3.4.2.3 Other variables derived from linked GP data

The following additional variables were created for the analysis of factors associated with participation in ALSPAC presented in Chapter 4.

3.4.2.3.1 Body mass index (BMI)

Read codes 22K.. (BMI), 229.. (O/E - height), and 22A.. (O/E - weight) were used to define BMI. Adolescent BMI was defined by calculating the mean of all measurements recorded from the age of 10 years. If there was only one measurement during this period, then this was the BMI value used.

3.4.2.3.2 Consultation rates

The GP data contains, for each individual, a series of dates, Read codes recorded on that date and – if relevant – a value associated with this code (e.g. weight in kg). Most codes are entered as a result of a consultation. However, a certain percentage are administration or other codes entered outside of a consultation. For example, the code 9N4C. (Failed encounter – no answer when rang back) would not be entered as part of a consultation. As previous researchers have done (Wang et al. 2013), I defined consultations by excluding any Read codes relating to administration, hospitalisations and the provision of services and by counting multiple consultations in a given day as one. For this thesis, I calculated consultation rates from 15-19 years; this was defined as the total number of consultations during this period, divided by five.

3.4.2.3.3 Drug counts

As in previous research (Brilleman and Salisbury 2012, Cornish et al. 2013) I counted the number of different drugs received by each individual at each year of age to provide an overall measure of morbidity. As described in the above papers, each unique drug was only counted once – so that repeat prescriptions and different formulations or doses of the same drug were not counted. In this thesis, I calculated the mean drug count from 15-19 years.

3.5 Summary statistics for all variables

This section gives summary statistics for all the variables used in this thesis, except the variables derived from the GP data (the latter are described in more detail in Chapters 6 and 7). Numbers and percentages are given for categorical variables and the mean and standard deviation or median and interquartile range (IQR) are given for numerical variables. Note that the denominators vary because the statistics given are for all singletons and twins alive at one year who had not withdrawn from the study for whom each variable was measured. Tables 3-4 and 3-5 describe the baseline covariates, Table 3-6 describes the exposures, Table 3-7 describes the outcomes, and Table 3-8 describes the auxiliary variables (included those obtained from the NPD but excluding those derived from GP data, as explained above).

Table 3-4: Summary statistics for maternal baseline covariates

Covariate	Level	N (%)[1]
Education	O level / lower	8,022 (65%)
	A level	2,791 (22%)
	Degree/higher	1,599 (13%)
Parity	0	5,771 (45%)
	1	4,539 (35%)
	2	1,849 (14%)
	3+	767 (6%)
Age (at birth of index child)	<20	650 (5%)
	20-24	2,688 (19%)
	25-29 (ref.)	5,404 (39%)
	30-34	3,849 (28%)
	35+	1,384 (10%)
Ethnicity	White	12,001 (97%)
	Non-white	323 (3%)
Age at first pregnancy	<20	2,631 (20%)
	20-24	4,031 (31%)
	25-29	4,507 (34%)
	30+	1,974 (15%)
Smoking (ever)	No	6,623 (51%)
	Yes	6,429 (49%)
Alcohol use in early pregnancy	No	10,176 (79%)
	Yes	2,641 (21%)
Drug use in early pregnancy	No	12,037 (97%)
	Yes	372 (3%)
Married	Yes	9,803 (75%)
	No	3,281 (25%)
Depression score (n=11,974)	Median (IQR)	6 (3-10)
Anxiety score (n=12,043)	Median (IQR)	4 (2-7)

1. Unless otherwise specified

Table 3-5: Summary statistics for child, paternal and family baseline covariates

Covariate	Level	N (%)[1]
Sex	Male	7,536 (51%)
	Female	7,148 (49%)
Education	O level / lower	6,662 (56%)
	A level	3,104 (26%)
	Degree/higher	2,168 (18%)
Smoking (ever)	No	4,419 (41%)
	Yes	6,271 (59%)
Depression score (n=9,699)	Median (IQR)	3 (1-6)
Anxiety score (n=9,652)	Median (IQR)	2 (1-4)
Family occupational social class	I-IIInm	9,269 (81%)
	IIIm-V	2,233 (19%)
Housing tenure	Mortgaged/owned	9,556 (73%)
	Private rented	933 (7%)
	Council/housing association	2,535 (19%)
Car use	Yes	11,624 (89%)
	No	1,407 (11%)
Phone in home	Yes	11,568 (89%)
	No/incoming only	1,503 (12%)
Double glazing	None	6,427 (52%)
	Full/partial	6,041 (48%)
Crowding index	≤0.5	5,329 (42%)
	>0.5 – 0.75	4,013 (31%)
	>0.75 – 1	2,579 (20%)
	>1	878 (7%)
Number of rooms (n=12,889)	Median (IQR)	5 (4-7)
Financial difficulties score (n=12,083)	Median (IQR)	2 (0-5)
Family adversity index (n=13,163)	Median (IQR)	1 (0-2)

1. Unless otherwise specified

Table 3-6: Summary statistics for exposure variables

Exemplar	Exposure	Level	N (%)[1]
1	Duration of breastfeeding	Never/<1 month	5,387 (43%)
		1 to <3 months	1,576 (13%)
		3 to <6 months	1,883 (15%)
		6 months+	3,719 (30%)
2	Smoking in pregnancy	No	9,861 (75%)
		Yes	3,294 (25%)
3	Teenage smoking (15 years):		
	Ever smoked	No	2,330 (44%)
		Yes	2,975 (56%)
	Frequency of current smoking	Never	4,410 (82%)
		< Daily	591 (11%)
		Daily	356 (7%)
	Cotinine	≤9.5ng/ml	3,146 (91%)
		>9.5ng/ml	307 (9%)

1. Unless otherwise specified

Table 3-7: Summary statistics for outcome variables

Exemplar	Outcome	Level	N (%)[1]
1	IQ at 15 years (n=4,951)	Mean (SD)	92 (13)
2	Depression at 18 years	No	4,203 (92%)
		Yes	360 (8%)
3	KS4 capped attainment score (n=12,020)	Mean (SD)	315 (96)
3	Obtained five or more A* to C grades at GCSE	Yes	7,212 (59%)
		No	4,934 (41%)

1. Unless otherwise specified

Table 3-8: Summary statistics for auxiliary variables not from GP data

Variable	Level	N (%)[1]
SEN status (Year 11) (n=11,395)	None	9,373 (82%)
	School action only	1,643 (14%)
	Statement	379 (3%)
% absence from school (Year 11) (n=11,395)	Median (IQR)	6 (3-11)
KS3 attainment score	Mean (SD)	226 (47)
KS2 attainment score	Mean (SD)	181 (45)
Emotional difficulties score (from SDQ at age 7) (n=8,204)	Median (IQR)	1 (0-2)

1. Unless otherwise specified

Chapter 4 Predictors of participation in ALSPAC

As described in the first two chapters of this thesis, the missing data methods that I use generally require the data to be MAR in order to be valid. Although there are some exceptions to this, the implication is that it is important to understand what factors are associated with missingness in a study – as this helps to determine what variables are needed in the analysis. In Chapter 2 I summarised previous findings regarding factors that predict participation in longitudinal studies in general as well as specifically in ALSPAC; this included two studies examining the association between genetic factors and participation (in ALSPAC). One limitation of the latter studies is that genetic data are only available on a (non-random) subsample of ALSPAC participants; genetic data are less likely to be available among individuals with lower socio-economic position, mothers with lower education levels and higher risk of depressive disorder, as well as those with higher scores on the Strengths and Difficulties Questionnaire (Martin et al. 2016).

In this chapter I describe the methods used to analyse participation (response) rates in ALSPAC over time and the predictors of these. I look at baseline predictors (socio-demographic and other factors) but also variables derived from the linked datasets, which are available for a large proportion of participants, regardless of participation in ALSPAC itself (in terms of questionnaire response and/or clinic attendance). I also analyse predictors of inclusion in the education and GP data. I present the results from these analyses and discuss their implications for the exemplars used throughout the thesis.

4.1 Predictors of continued participation in ALSPAC

This analysis was based on the 13,972 singletons and twins who enrolled during pregnancy (since those enrolling at age 7 or subsequently would not have baseline data available), were alive at 1 year, and had not subsequently withdrawn from the study. For the analysis of predictors of continued participation derived from GP data, individuals for whom ALSPAC had no NHS ID number (n=108) were excluded, leaving 13,864 (99% of the original 13,972) individuals. Among these, 2,777 had no GP data available, leaving n=11,087 who could potentially be included in the analysis.

4.1.1 Methods

For each ALSPAC questionnaire and study clinic, a binary variable was created to indicate whether or not an individual returned that questionnaire or attended the clinic. The questionnaires were grouped according to whether they were mother/carer-completed (these included both questionnaires about themselves and questionnaires about their child and will be referred to as mother-completed in the remainder of this chapter) or child-completed. Study clinics were grouped with the child-completed questionnaires. Arguably attendance at these would be largely determined by their parent(s), particularly at the younger ages; however, since the same might be said for the earlier child-completed questionnaires, it was felt that this was the most logical classification.

A wide range of baseline socio-demographic and other variables potentially associated with participation were included in the initial analysis. The majority of these variables were measured in pregnancy since this was when response rates were highest. However, since it was known from previous work that duration of breastfeeding was strongly related to study participation, this was also included. Paternal factors were not included, partly because response rates for these were lower but also because, in a preliminary analysis, none of the paternal factors considered (paternal education, smoking, and depression score) were associated with

participation after taking account of maternal factors. The baseline variables included are listed below; further details about these were given in Chapter 3 (Section 3.3.3).

Table 4-1: List of baseline variables included in the analysis of participation

	Variable[1]	Further information/categories
Maternal factors	Maternal education	O level/lower, A level, degree/higher
	Maternal age	<20, 20-24, 25-29, 30-34, 35+ years
	Parity	0, 1, 2+
	Age at first birth	<20, 20-24, 25-29, 30+ years
	EPDS depression score	Range 0-30
	Smoking in pregnancy Ever smoked?	
	Ethnicity	White/non-white
	Marital status	Married vs single/widowed/divorced/separated
	Duration of breastfeeding	Never/<1 month, 1-2 months, 3-5 months, 6+ months
Family / combined parental factors	Occupational social class	I-IIInm vs IIIm-V
	Housing tenure	Owned/mortgaged, private rented, other
	Number of rooms in house	Numerical
	Crowding index	Number of people/number of rooms: ≤0.5, 0.51-0.75, 0.751-1, >1
	Telephone in house	Yes, no/incoming only
	Use of car	By mother/carer or partner
	Double glazing	Full/partial, none
	Financial difficulties	Range 0-15, with higher score indicating greater financial difficulties
	Adversity index	Range 0-11, with higher score indicating greater adversity; categorised: 0, 1, 2, 3+
Child factors	Sex	

1. These variables are described in greater detail in Chapter 3 (Section 3.3.3)

Two separate random effects logistic regression models (one for child-completed participation and one for mother-completed participation) were used to model participation over time. Cubic splines (Durrleman and Simon 1989) with five knots were used to model the effects of age (for children) and time in study (for mothers). The knots were placed using Stata's default method recommended by Harrell (Harrell

2001). Although individuals were nominally sent questionnaires and invited to clinics at a given age, there was inevitably some variability in terms of the actual age at completion/attendance. However, for the purposes of this analysis, the (fixed) age at which the questionnaire/clinic invitation was sent, rather than the actual age at completion/attendance was used. For the mother-completed questionnaires, time in study was used instead of age since some questionnaires were completed in pregnancy. Thus, time in study = 0 was used to denote the beginning of pregnancy.

Firstly, all baseline variables were included in the models. Then, among a subset of correlated measures of socio-economic position (number of rooms in the house, crowding index, family adversity index, financial difficulties score, housing tenure, double glazing, use of car and phone) those that were not associated with participation (based on p-values: $p>0.3$) were removed before assessing whether the variables derived from linked datasets were associated with participation after taking account of baseline factors. This strategy was used in order to maximise the number of individuals included in the second stage of this analysis and to avoid the inclusion of highly correlated indicators of socio-economic position. The linked variables examined in relation to study participation are listed in Table 4-2.

Interaction terms were used to examine whether the impact of the different factors on child-completed participation changed over time. Since there was evidence for an interaction with time (age) for many of the factors, separate models were used to examine associations within the following three age bands: up to age 11 years, 11-15 years, and 16 years and over. I focussed only on child participation because the outcomes in Exemplars 1 and 3 (IQ and depression at 18) and the exposure in Exemplar 3 (teenage smoking) were all measured during study clinics (defined as child participation in this thesis).

Table 4-2: Variables from linked datasets using in the analysis of participation

Dataset	Variable[1]	Notes
NPD	KS4 attainment score	Range 0-540
	% absence in Year 11 (age 15-16)	
	SEN status in Year 11	No SEN, school support, statement of SEN
	KS3 attainment score	Range 0-430
	KS2 attainment score	Range 0-280
GP data	Smoking record by age 18	Yes/no
	BMI	Continuous
	Depression before age 18	Yes/no
	Consultation rate age 15-19	≤1 per year, >1–4 per year, >4 per year
	Prescription rate age 15-19	≤1 per year, >1–4 per year, >4 per year

1. These variables are described in more detail in Chapter 3 (Section 3.4.1 and 3.4.2)

4.1.2 Results

Figure 4-1 gives response rates across time for the child-completed and mother-completed questionnaires/clinics, respectively. As shown in the figure, participation rates were high in pregnancy and gradually dropped off during early childhood; they then remained relatively stable for a while, dropping again during mid-late adolescence.

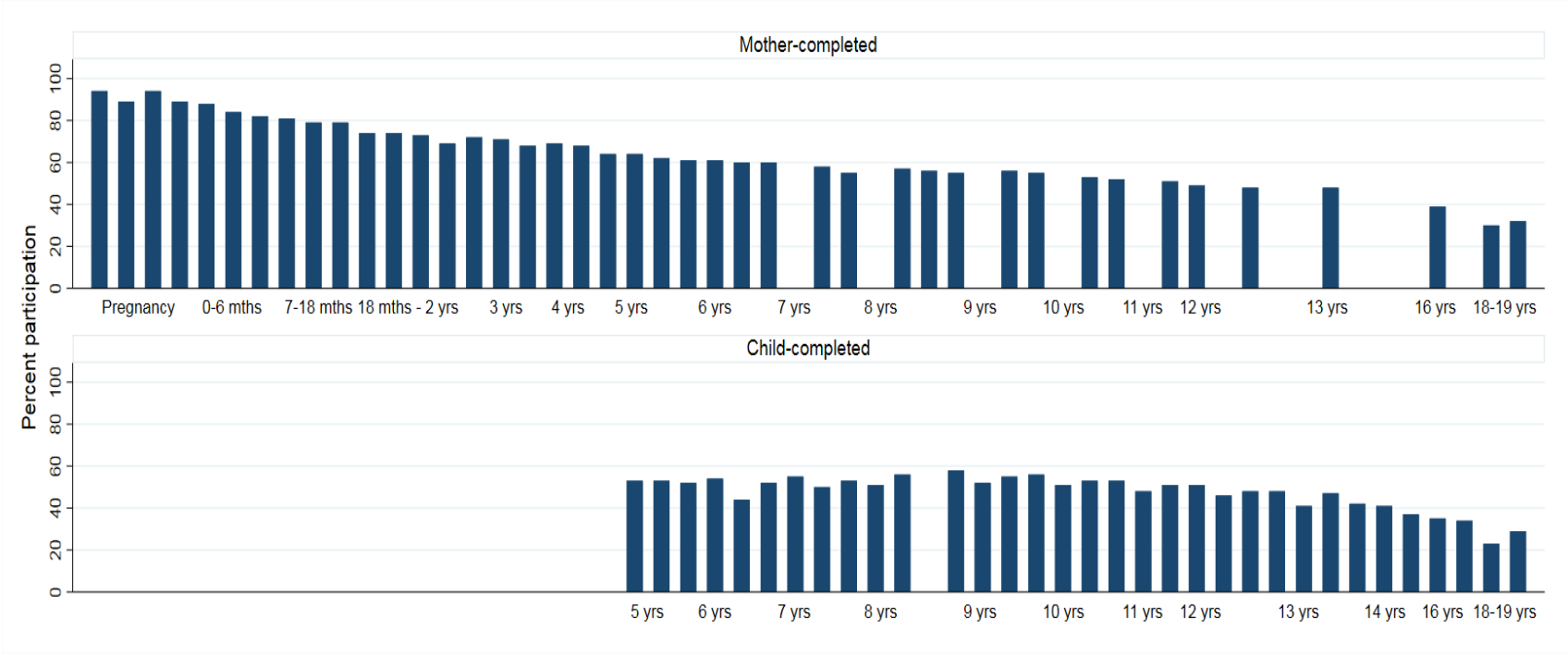


Figure 4-1: Participation rates (%) in ALSPAC: mother and child-completed

Table 4-3: Odds ratios for participation for child and maternal baseline covariates (n=9,049)

Covariate	Level	Child participation		Mother participation	
		OR (95% CI)[1,2]	p-value	OR (95% CI)[1,2]	p-value
Sex	Female vs male	1.86 (1.67, 2.06)	<0.001	1.06 (0.94, 1.20)	0.3
Mother's education	O level / lower	1.00		1.00	
	A level	1.48 (1.30, 1.69)		1.66 (1.43, 1.94)	
	Degree/higher	1.75 (1.46, 2.08)	<0.001	1.99 (1.62, 2.44)	<0.001
Parity	0	1.00		1.00	
	1	0.77 (0.66, 0.89)		0.76 (0.64, 0.90)	
	2	0.60 (0.48, 0.76)		0.59 (0.45, 0.77)	
	3+	0.47 (0.33, 0.66)	<0.001	0.53 (0.36, 0.78)	<0.001
Mother's age (at birth of index child)	<20	0.35 (0.24, 0.53)		0.40 (0.25, 0.62)	
	20-24	0.65 (0.54, 0.78)		0.62 (0.51, 0.77)	
	25-29 (ref.)	1.00		1.00	
	30-34	1.45 (1.25, 1.68)		1.55 (1.30, 1.85)	
	35+	1.75 (1.40, 2.19)	<0.001	1.98 (1.53, 2.55)	<0.001
Mother's ethnicity	Non-white vs white	0.54 (0.37, 0.79)	0.002	0.29 (0.19, 0.45)	<0.001
Age at first pregnancy	<20	1.00		1.00	
	20-24	1.30 (1.09, 1.56)		1.35 (1.10, 1.66)	
	25-29	1.34 (1.10, 1.64)		1.63 (1.30, 2.05)	
	30+	1.42 (1.10, 1.84)	0.02	1.75 (1.30, 2.36)	<0.001
Maternal smoking	Yes vs no (in pregnancy)	0.80 (0.68, 0.94)	0.006	0.77 (0.65, 0.93)	0.006
Duration of breastfeeding	Yes vs no (ever)	0.78 (0.69, 0.89)	<0.001	0.82 (0.71, 0.95)	<0.001
	Never/<1 month	1.00		1.00	
	1 to <3 months	1.70 (1.44, 2.01)		1.67 (1.37, 2.02)	
	3 to <6 months	1.79 (1.53, 2.10)		1.72 (1.44, 2.07)	
	6 months+	2.24 (1.95, 2.57)	<0.001	2.38 (2.03, 2.79)	<0.001
Married	Yes vs no	1.05 (0.90, 1.22)	0.6	1.08 (0.91, 1.28)	0.4
Depression score	Per 1 unit increase in score	0.99 (0.98, 1.01)	0.3	0.98 (0.97, 1.00)	0.03

1. Random intercepts model with cubic splines for age.
2. Also adjusted for other (family) covariates presented in Table 4-4.

Table 4-4: Odds ratios for participation for other baseline covariates (n=9,049)

Covariate	Level	Child participation		Mother participation	
		OR (95% CI)[1,2]	p-value	OR (95% CI)[1,2]	p-value
Family social class	Manual vs non-manual	0.84 (0.72, 0.97)	0.02	0.72 (0.61, 0.86)	<0.001
Housing tenure	Owned/mortgaged	1.00		1.00	
	Private rented	0.62 (0.48, 0.79)		0.57 (0.43, 0.76)	
	Council/HA/other	0.86 (0.71, 1.04)	<0.001	0.82 (0.66, 1.03)	<0.001
Number of rooms[3]	Per 1 room increase	1.05 (0.99, 1.11)	0.09	1.02 (0.96, 1.09)	0.5
Phone in home	Yes vs no/incoming	0.70 (0.56, 0.88)	0.002	0.70 (0.54, 0.91)	0.008
Car use	No vs yes	0.70 (0.55, 0.89)	0.003	0.76 (0.58, 1.00)	0.05
Double glazing	None vs full/partial	0.88 (0.79, 0.98)	0.02	0.86 (0.76, 0.98)	0.02
Family adversity index[3]	0	1.00		1.00	
	1	0.92 (0.81, 1.05)		0.84 (0.72, 0.98)	
	2	0.86 (0.72, 1.03)		0.83 (0.67, 1.01)	
	3+	0.88 (0.71, 1.08)	0.4	0.88 (0.70, 1.12)	0.1
Financial difficulties[3]	Per 1 unit increase	0.98 (0.97, 1.00)	0.08	0.99 (0.97, 1.01)	0.07
Crowding index[3]	≤0.5	1.00		1.00	
	>0.5 – 0.75	0.95 (0.81, 1.11)		0.84 (0.70, 1.01)	
	>0.75 – 1	0.86 (0.70, 1.07)		0.78 (0.61, 1.00)	
	>1	0.77 (0.55, 1.07)	0.4	0.61 (0.42, 0.89)	0.06

1. Random intercepts model with cubic splines for age
2. Also adjusted for child and maternal covariates presented in Table 4-3
3. Family adversity index and crowding were omitted from subsequent analyses of child participation and number of rooms in the house was omitted from mother participation

4.1.2.1 Baseline predictors of participation

Table 4-3 and 4-4 give mutually adjusted odds ratios for child-completed and mother-completed participation for all the baseline covariates considered. On the whole, the odds ratios for child-completed participation were similar to those for mother-completed participation. There were two main exceptions to this: firstly, sex was a strong predictor of child-completed but not mother-completed participation; secondly, maternal ethnicity was more strongly related to mother-completed participation than child-completed participation.

Table 1 in Appendix B shows the odds ratios for child participation in the three age bands for the baseline covariates. Sex became more strongly associated with participation (by the child) with age: the odds ratios were 1.62 (95% CI: 1.45, 1.80) under 11 years, 2.50 (2.16, 2.89) for 11-15 years, and 3.61 (3.18, 4.09) for 16 years and over. Conversely, the effect of maternal ethnicity became weaker with age

(OR=0.47, 0.50 and 0.69 at <11 years, 11-15 years and 16+ years, respectively); similarly, marital status (of the mother) was more strongly associated with participation up to 11 years, whereas there was no evidence that it was associated with participation at ages 11-15 or 16+ years after taking account of the other factors. Many of the other baseline factors (parity, mother's age at birth of child, smoking in pregnancy, duration of breastfeeding, number of rooms in home, car and phone ownership) were more strongly associated with participation between 11 and 15 years than at the other ages (Appendix B, Table 1).

4.1.2.2 Predictors of participation: offspring education variables

Odds ratios for the five NPD variables (adjusted for baseline factors) for both child-completed and mother-completed participation are given in Table 4-5. All three factors (attainment, absence and SEN status) were associated with child-completed participation; attainment and absence (but not SEN status) were associated with mother-completed participation. Of the attainment scores, KS2 attainment was most strongly associated with participation. After adding the education variables into the model, some of the associations between the (maternal but not other) baseline covariates and participation were slightly weakened, particularly mother's education, ethnicity, and breastfeeding duration; however, the changes were relatively small and these factors remained strongly associated with participation (results not shown).

Table 4-5: Odds ratios for participation: linked education (NPD) variables (n=6,136)

		Child participation		Mother participation	
Factor	Level	OR (95% CI)[1,2]	p-value	OR (95% CI)[1,2]	p-value
KS4 attainment	For 10 point increase	1.04 (1.02, 1.05)	<0.001	1.03 (1.01, 1.04)	0.002
SEN status (year 11)	None	1.00		1.00	
	School action	0.83 (0.67, 1.02)		0.99 (0.77, 1.27)	
	Statement	0.71 (0.37, 1.38)	0.1	1.05 (0.48, 2.31)	>0.9
School absence (year 11)	For 1 pt increase in square root of % absence	0.85 (0.81, 0.90)	<0.001	0.83 (0.78, 0.88)	<0.001
KS3 attainment	For 10 point increase	1.02 (1.00, 1.04)	0.04	1.02 (0.99, 1.04)	0.1
KS2 attainment	For 10 point increase	1.05 (1.03, 1.08)	<0.001	1.06 (1.03, 1.08)	<0.001

1. Random intercepts model with cubic splines for age
2. Adjusted for baseline factors

Table 4-6 gives odds ratios for child participation for the linked education variables for the three different age bands. The association between attainment at Key Stage 4 (age 16 years) and participation became stronger with age, whereas the association with Key Stage 2 attainment (age 11 years) became weaker. School absence in Year 11 (age 15-16) and Key Stage 3 attainment (age 14 years) was more strongly associated with participation at 11-15 and 16+ years than participation up to 11 years. Finally, special education needs in Year 11 was more strongly associated with participation at <11 and 11-15 years than at 16 years and above.

Table 4-6: Odds ratios for child participation at different ages: linked education variables (n=6,136)

Factor	Level	OR (95% CI)[1,2]		
		< 11 years	11-15 years	16+ years
KS4 attainment	For 10 point increase	1.03 (1.01, 1.04)	1.06 (1.04, 1.08)	1.10 (1.08, 1.12)
SEN status (year 11)	None	1.00	1.00	1.00
	School action	0.80 (0.65, 0.99)	0.77 (0.60, 1.04)	0.97 (0.75, 1.27)
	Statement	0.58 (0.29, 1.14)	0.61 (0.23, 1.57)	1.45 (0.63, 3.33)
School absence (year 11)	For 1 pt increase in sq. root of % absence	0.87 (0.83, 0.92)	0.78 (0.73, 0.84)	0.80 (0.75, 0.86)
KS3 attainment	For 10 point increase	1.01 (0.99, 1.03)	1.03 (1.01, 1.06)	1.03 (1.01, 1.05)
KS2 attainment	For 10 point increase	1.07 (1.04, 1.09)	1.06 (1.02, 1.09)	1.03 (1.00, 1.05)

1. Random intercepts model with cubic splines for age
2. Adjusted for baseline factors

4.1.2.3 Predictors of participation: variables from GP data

Odds ratios for both child and mother-completed participation for the GP-derived measures are given in Table 4-7. After adjusting for baseline factors associated with participation, all the factors considered were strongly related to child participation in ALSPAC and most of them appeared to be associated with mother participation, although the associations were generally – but not exclusively - weaker than those for child participation.

Table 4-7: Odds ratios for participation: GP-derived offspring measures

		Child participation		Mother participation	
Factor	Level	OR (95% CI)[1,2]	p-value	OR (95% CI)[1,2]	p-value
Smoking record by age 18[3a]	Yes v s no	0.64 (0.53, 0.76)	<0.001	0.68 (0.55, 0.84)	<0.001
Depression before age 18[3b]	Yes vs no	0.71 (0.56, 0.90)	0.005	0.80 (0.60, 1.07)	0.1
BMI[3c]	per 1kg/m^2	0.98 (0.96, 0.99)	0.001	0.97 (0.95, 0.99)	0.001
Consultation rate age 15-19[3d]	≤1 per year	1.00		1.00	
	>1 – 4 per year	1.49 (1.26, 1.76)		1.23 (1.01, 1.50)	
	>4 per year	1.75 (1.46, 2.11)	<0.001	1.26 (1.01, 1.57)	0.08
Prescription rate age 15-19[3d]	≤1 per year	1.00		1.00	
	>1 – 4 per year	1.41 (1.23, 1.62)		1.19 (1.01, 1.41)	
	>4 per year	1.53 (1.25, 1.87)	<0.001	1.20 (0.94, 1.53)	0.09

1. Random intercepts model with cubic splines for age
2. Adjusted for baseline factors
3. a) n= 5,527 & 5,513; b) n=5,413 & 5,399; c) n=4,290 & 4,280; d) n=5,477 & 5,464 for child and mother participation, respectively.

All the GP measures were more strongly associated with child participation at ages 11-15 years than at the other ages. The odds ratios for consultation rates and prescription rates at ages 15-19 were similar for participation aged <11 years and participation aged 16 years and over, whereas the association between BMI and depression was weaker for participation at ages 16 and above (and the 95% confidence intervals included the null) than for participation under 11 years. These results are shown in Table 4-8.

Table 4-8: Odds ratios for child participation at different ages: linked GP-derived offspring measures

Factor	Level	OR (95% CI)[1,2]		
		< 11 years	11-15 years	16+ years
Smoking record by age 18	Yes v s no	0.73 (0.60, 0.89)	0.51 (0.40, 0.60)	0.62 (0.52, 0.73)
Depression before age 18	Yes vs no	0.66 (0.50, 0.87)	0.64 (0.45, 0.90)	0.81 (0.65, 1.01)
BMI	per 1kg/m^2	0.97 (0.96, 0.99)	0.96 (0.94, 0.98)	0.99 (0.98, 1.01)
Consultation rate age 15-19	≤1 per year	1.00	1.00	1.00
	>1 – 4 per year	1.44 (1.20, 1.74)	1.78 (1.41, 2.25)	1.42 (1.22, 1.65)
	>4 per year	1.58 (1.29, 1.94)	2.31 (1.78, 3.00)	1.73 (1.46, 2.05)
Prescription rate age 15-19	≤1 per year	1.00	1.00	1.00
	>1 – 4 per year	1.36 (1.16, 1.58)	1.66 (1.36, 2.01)	1.28 (1.13, 1.45)
	>4 per year	1.39 (1.11, 1.74)	1.90 (1.44, 2.53)	1.43 (1.19, 1.71)

1. Random intercepts model with cubic splines for age
2. Adjusted for baseline factors

4.2 Predictors of inclusion in the education data

This analysis is based on all singletons and twins who enrolled during pregnancy (in order to restrict to those with baseline covariates), were alive at one year and had not subsequently withdrawn from the study (n=13,975). Of these individuals, 12,038 (86%) had at least one linked education variable (i.e. were included in the linked education data).

4.2.1 Methods

The outcome for this analysis was binary: any education data (yes/no), so was equal to 1 (linked education data) for 12,038 individuals and equal to zero (no linked education data) for 1,937 individuals. Logistic regression was used to analyse predictors of this outcome. All the baseline socio-demographic variables included in the analysis of participation in ALSPAC were included as predictors; in addition, father's education was also included. Again, among a subset of correlated measures of socio-economic position (number of rooms in the house, crowding index, family adversity index, financial difficulties score, housing tenure, double glazing, use of car and phone) those that were not associated with inclusion in the education data (based on p-values as well as estimates of odds ratios and their confidence intervals)

were removed in order to retain the maximum sized dataset (n=9,186) for this analysis.

4.2.2 Results

Individuals whose parents were more highly educated were less likely to have linked education data, as were males, those breastfed for longer, and first-born children. Missingness in the linked education data (having no linked education data at all) was also associated with other indicators of socio-economic position – generally, although not exclusively, a higher odds of missingness was associated with higher socio-economic position (Table 4-9).

Table 4-9: Predictors of inclusion in the linked education data among individuals with complete baseline covariates (n=9,186)

Factor	Level	OR (95% CI)	p-value
Mother's education	O level or lower	1.00	
	A level	0.70 (0.60, 0.83)	
	Degree or higher	0.51 (0.42, 0.63)	<0.001
Father's education	O level or lower	1.00	
	A level	0.72 (0.61, 0.84)	
	Degree or higher	0.55 (0.45, 0.66)	<0.001
Sex	Female vs male	1.24 (1.09, 1.40)	0.001
Parity	0	1.00	
	1	1.10 (0.95, 1.28)	
	2	1.24 (0.99, 1.56)	
	3+	1.48 (1.02, 2.16)	0.1
Duration of breastfeeding	Never/<1 month	1.00	
	1 to <3 months	0.88 (0.71, 1.09)	
	2 to <6 months	0.70 (0.58, 0.85)	
	6+ months	0.83 (0.70, 0.98)	0.004
Car use	Yes vs no	0.71 (0.53, 0.94)	0.02
Number of rooms	Per 1 room increase	0.95 (0.91, 1.01)	0.1
Maternal smoking	Never	1.00	
	Yes, not in pregnancy	0.93 (0.80, 1.08)	
	In pregnancy	0.81 (0.67, 0.97)	0.07
Housing tenure	Mortgaged/owned	1.00	
	Private rented	0.61 (0.47, 0.78)	
	Council/HA/other	0.85 (0.67, 1.08)	<0.001
Mother's ethnicity	Non-white vs white	0.70 (0.47, 1.04)	0.08
Family social class	Manual vs non-manual	1.15 (0.93, 1.42)	0.2
Age at first pregnancy	<20	1.00	
	20-24	1.08 (0.86, 1.36)	
	25-29	1.25 (0.98, 1.59)	
	30+	1.21 (0.90, 1.62)	0.3
Married	Yes vs no	1.11 (0.92, 1.33)	0.3
Mother's age (at birth of index child)	<20	1.00	
	20-24	0.73 (0.43, 1.25)	
	25-29	0.87 (0.51, 1.51)	
	30-34	0.78 (0.44, 1.38)	
	35+	0.73 (0.40, 1.33)	0.3

4.3 Predictors of inclusion in the GP data

As above (Section 4.2), this analysis is based all singletons and twins who enrolled during pregnancy, were alive at one year and had not subsequently withdrawn from the study. As when looking at the association between GP measures and ALSPAC participation, it is additionally restricted to those for whom ALSPAC had an NHS ID number (n=13,864). As explained in Chapter 3, there were a variety of reasons why someone would not have any linked GP data. Of the 13,864 individuals included in this analysis, a total of 2,777 (20%) had no linked GP data. Among these, 1,538 (55%) were not sent fair processing materials and 364 (13%) explicitly dissented to linkage to their health data. For the remaining 875, the most likely explanation is that they moved out of the area before registering with a GP, since the vast majority of the linked GP data comes from local GP practices.

4.3.1 Methods

The outcome for this analysis was binary: linkage to any GP data (yes/no), so was equal to 1 (linked GP data) for n=13,864-2,777=11,087 individuals and equal to zero (no linked GP data) for 2,777 individuals. Logistic regression was used to analyse predictors of this outcome. All the baseline socio-demographic variables included in the analysis of participation in ALSPAC were included as predictors; in addition, father's education was also included. As above, among a subset of correlated measures of socio-economic position (number of rooms in the house, crowding index, family adversity index, financial difficulties score, housing tenure, double glazing, use of car and phone) those that were not associated with inclusion in the GP data were removed in order to retain the maximum sized dataset (n=9,095) for this analysis.

4.3.2 Results

Table 4-10 shows odds ratios for having any linked GP data. Males were less likely to have linked GP data, as were children with more educated fathers, those living in private rented accommodation, those whose mother smoking during pregnancy, those who were breastfed for longer, those whose mother was older at their first

pregnancy, those whose mother was not married, and those whose family occupational social class was classified as non-manual.

Table 4-10: Predictors of non-inclusion in GP extract (n=9,095 with baseline covariates)

Factor	Level	No GP data OR (95% CI)	p-value
Sex	Female vs male	0.84 (0.75, 0.93)	0.001
Mother's education	O level / lower	1.00	
	A level	1.08 (0.94, 1.21)	
	Degree/higher	1.06 (0.88, 1.27)	0.5
Parity	0	1.00	
	1	1.12 (0.98, 1.27)	
	2	1.19 (0.98, 1.43)	
	3+	1.29 (0.97, 1.72)	0.2
Mother's age (at birth of index child)	<20	1.00	
	20-24	1.24 (0.80, 1.94)	
	25-29	1.13 (0.72, 1.77)	
	30-34	1.06 (0.66, 1.71)	
	35+	1.19 (0.72, 1.96)	0.5
Mother's ethnicity	Non-white vs white	1.14 (0.78, 1.68)	0.5
Family social class	Manual vs non-manual	0.87 (0.74, 1.02)	0.09
Age at first pregnancy	<20	1.00	
	20-24	1.20 (0.99, 1.45)	
	25-29	1.37 (1.11, 1.69)	
	30+	1.49 (1.15, 1.95)	0.02
Maternal smoking	Never	1.00	
	Yes, not in pregnancy	0.98 (0.86, 1.12)	
	In pregnancy	1.16 (1.01, 1.34)	0.07
Duration of breastfeeding	Never/<1 month	1.00	
	1 to <3 months	1.16 (0.97, 1.39)	
	3 to <6 months	1.15 (0.98, 1.35)	
	6 months+	1.21 (1.05, 1.39)	0.05
Married	Yes vs no	0.88 (0.76, 1.03)	0.1
Housing tenure	Owned/mortgaged	1.00	
	Private rented	1.71 (1.39, 2.12)	
	Council/HA rented/other	1.31 (1.09, 1.56)	<0.001
Father's education	O level / lower	1.00	
	A level	1.13 (0.99, 1.29)	
	Degree / higher	1.51 (1.28, 1.79)	<0.001

4.4 Implications for exemplars

A large proportion of the baseline factors that are associated with participation in ALSPAC from this analysis are markers of socio-economic position. As mentioned in Chapter 3 (Section 3.1.1), maternal age and paternal occupational social class have also been shown to be associated with initial participation in ALSPAC. Since the associations being considered in all three exemplars included in this thesis (breastfeeding and IQ, maternal smoking and offspring depression, teenage smoking and educational attainment) are likely to be confounded by social position, it will be important to include all of the factors associated with initial and child-based participation as covariates in the complete case analysis for each exemplar.

The analysis above also suggests that there could be some unmeasured causes of non-response in ALSPAC. For example, school absence at 15-16 years and GP consultation rates aged 15-19 were associated with non-response after conditioning on a wide range of baseline covariates. However, these cannot be a cause of participation in ALSPAC before age 15, which suggests that they must be associated with additional (perhaps unmeasured) predictors of non-response. Further, these – or other – unmeasured factors may also be associated with initial participation in ALSPAC. This indicates that the data for any given analysis are unlikely to be MAR conditional on the observed variables, and thus that it will be important to assess sensitivity to MNAR in all exemplars.

Rather than considering analyses in general, however, I will now focus specifically on the analyses being carried out in the next three chapters of this thesis.

4.4.1 Exemplar 1: Breastfeeding and IQ

Figure 4-2 shows the DAG (directed acyclic graph) that I hypothesise could represent my first exemplar. In this and the other DAGs drawn below, C represented measured covariates, U represents unmeasured factors and R represents missingness (R=1 if the individual is a complete case; R=0 otherwise), and SEP represents various markers of socio-economic position (occupational social class, housing, use of car and phone,

financial difficulties, number of rooms in home, crowding index, and family adversity index). In this exemplar I hypothesise that the exposure (breastfeeding), all the measured covariates listed in the figure and the outcome (IQ) are associated with missingness. [Note that SEN status is variable indicating whether an individual was classified as having SEN in Year 11. This is a marker of whether an individual has one or more conditions (which could impact on a range of physical, cognitive, and/or behavioural factors). Thus, I am regarding SEN status as being caused by these underlying conditions through these physical, cognitive, and behavioural factors.] My hypothesis is that attainment is only associated with missingness through its association with IQ but that school absence and SEN status could be associated with missingness both through their association with IQ and through an association with unmeasured factors that may also confound the association between breastfeeding and IQ. The analysis carried out above supports this DAG – I showed above that breastfeeding and attainment were both associated with missingness, as were SEN status, and school absence. This suggests that the complete case analysis will be biased in this situation – since missingness is dependent on the outcome, conditional on the variables included in the analysis model (breastfeeding, C and IQ). Most of the covariates listed in the figure below were hypothesised to be confounders of the breastfeeding-IQ association. Ethnicity and age at first pregnancy were predictors of missingness and also hypothesised to be predictors of breastfeeding but not IQ. There may also be unmeasured covariates that are associated with missingness.

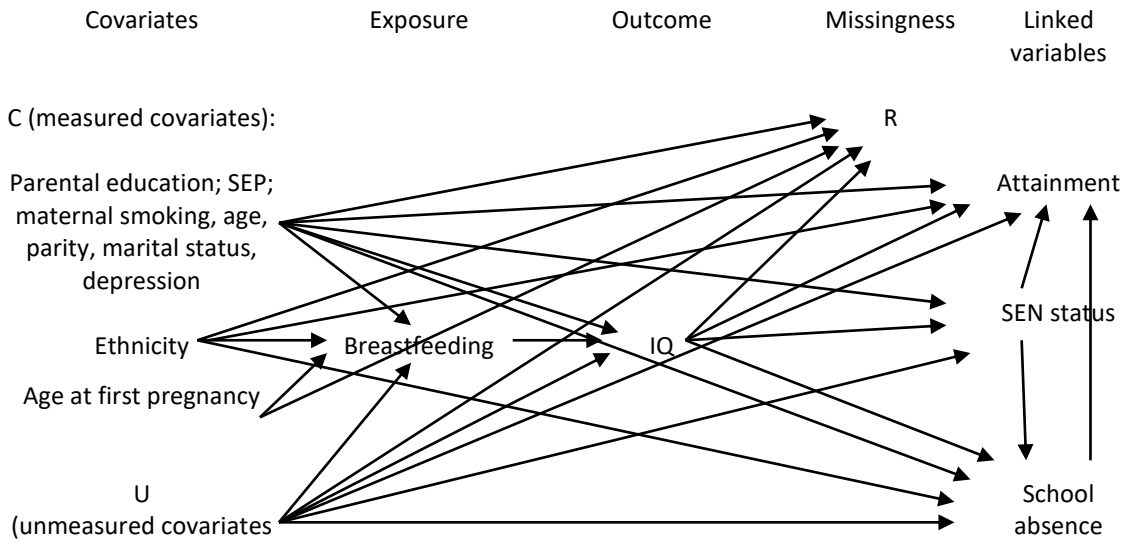


Figure 4-2: Hypothesised DAG for Exemplar 1: breastfeeding and IQ

If attainment, SEN status and school absence were used as auxiliary variables in MI, this should reduce bias because – in the imputation model – this will give a better approximation to MAR (assuming these variables have a reasonably high joint correlation with IQ). I can include attainment at various ages to further improve the estimation of IQ. The inclusion of school absence should also help by providing a proxy of the unmeasured predictors of missingness; again, this will result in a closer approximation to MAR in the imputation model.

4.4.2 Exemplar 2: Smoking in pregnancy and offspring depression

The DAG for this exemplar is shown in Figure 4-3. As above, I hypothesise that the exposure (smoking in pregnancy), many (but not all) of the measured baseline covariates, and the outcome (offspring depression) are associated with missingness. Most of the measured covariates listed in Figure 4-3 were also hypothesised to be confounders of the smoking in pregnancy – offspring depression association. My hypothesis is that GP-recorded depression is associated with missingness only through its association with offspring depression. Although the hypothesis is that missingness depends on both the outcome and exposure, because offspring depression is a binary variable and logistic regression will be used in this analysis, the

exposure odds ratio obtained from the complete case analysis could be unbiased (if there is no multiplicative interaction (on the probability scale) between smoking in pregnancy and offspring depression with respect to missingness). However, the latter also assumes that missingness is truly associated with the binary outcome (depression: yes/no) rather than an underlying continuous measure (i.e. depression severity), which seems unlikely. In this case, it is unclear how much bias might be introduced in the complete case analysis.

If multiple imputation were used to impute missing depression, one of the key covariates in the imputation model for offspring depression would be GP-recorded depression. I have shown that missingness in this variable is associated with several socio-demographic variables. There could also be one or more unmeasured predictors of missingness in GP-recorded depression, which implies that this variable could be MNAR. In this imputation model (for missing offspring depression), it is likely that missingness is less dependent on the outcome (conditional on GP-recorded depression and other measured covariates), although this would depend on how strongly associated GP-recorded depression is with ALSPAC-recorded depression. As such, we might expect bias to be reduced by using multiple imputation.

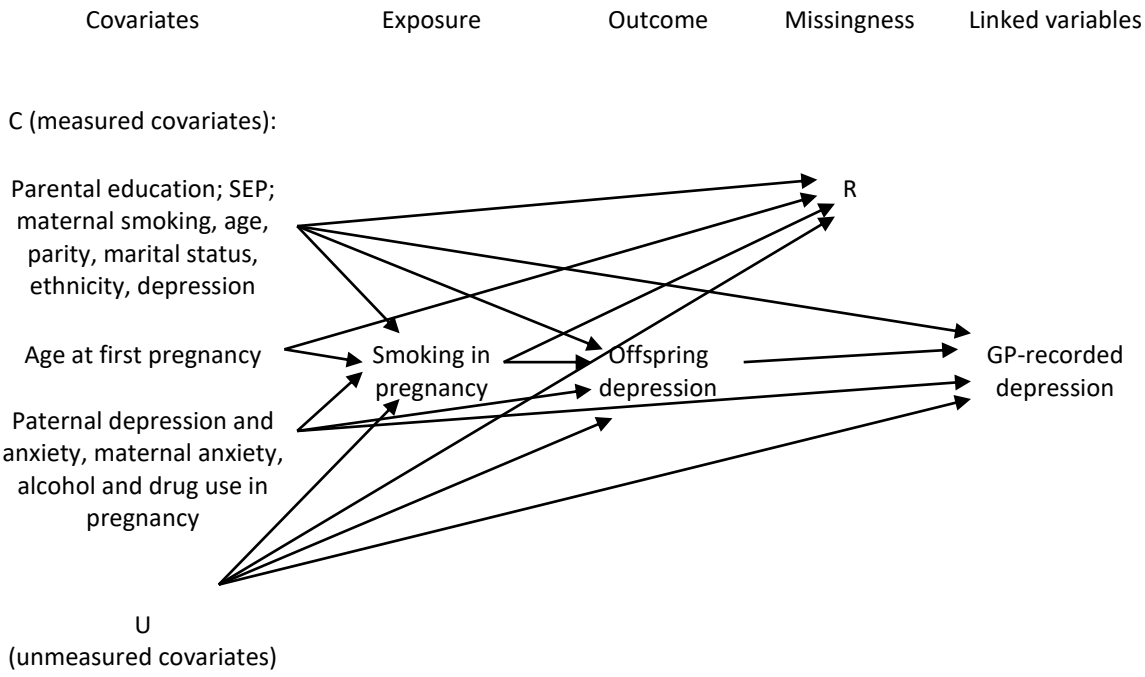


Figure 4-3: Hypothesised DAG for Exemplar 2: smoking in pregnancy and offspring depression

4.4.3 Exemplar 3: Teenage smoking and educational attainment

Finally, Figure 4-4 represents the hypothesised DAG for Exemplar 3. In this we expect the exposure (teenage smoking) to be MNAR conditional on baseline socio-demographic covariates (C). We also know that educational attainment is MNAR because the main reason for having missing educational attainment is attendance at an independent school (which, in general, is associated with higher attainment). Thus, we hypothesis that missingness depends on both the exposure and the outcome (as well as other covariates). There are two outcomes in this analysis: the capped GCSE point score, a continuous variable, and whether or not a person obtained five or more A* to C grades at GCSE, a binary variable. Thus, the complete case analysis of the association between smoking and GCSE score will be biased, whereas the complete case analysis of the association between smoking and the binary outcome (if analysed using logistic regression) could be unbiased if there is no multiplicative interaction between smoking and obtaining five or more A* to C grades with respect to the probability of being a complete case. However, the latter also assumes that

missingness is truly associated with the binary outcome rather than the underlying continuous attainment score, which seems unlikely. Again, it is unclear how much bias might be introduced in this situation. In both cases we would expect multiple imputation to produce biased estimates of the exposure-outcome association. This bias could potentially be reduced by including GP-recorded smoking as an auxiliary variable in the imputation models.

Note that other teenage behavioural factors (in particular, alcohol consumption and other risk-taking behaviours) are likely to be associated with teenage smoking and also causally associated with educational attainment. However, my hypothesis is that teenage smoking and these other behavioural factors share common causes (socio-demographic and parental factors); thus these common causes form a sufficient set of necessary confounders to adjust for. For this reason, these additional behavioural factors are not included in Figure 4-4.

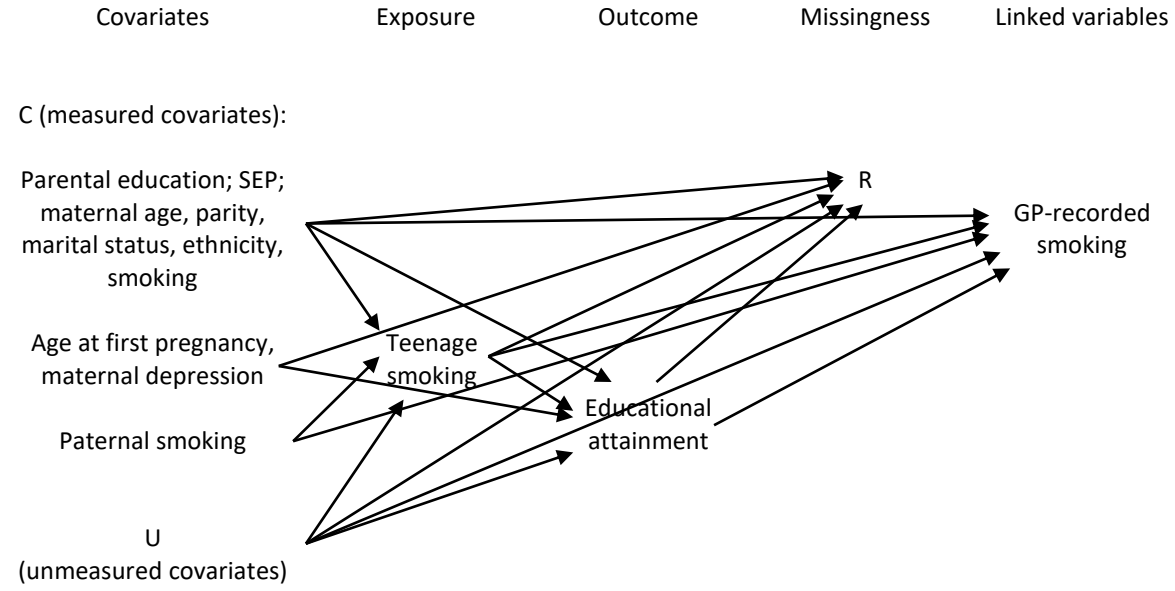


Figure 4-4: Hypothesised DAG for Exemplar 3: teenage smoking and educational attainment

Chapter 5 Missing continuous outcome

This and the following two chapters further address objective 1 (to use linked health and administrative data to examine patterns and predictors of missing data in ALSPAC). In particular, I use the linked variables to examine whether specific outcomes (Chapters 5 and 6) or exposures (Chapter 7) are likely to be MNAR. This and the following two chapters also address objective 2 (to incorporate linked health and administrative data as auxiliary variables in multiple imputation and other models to examine bias in estimates of exposure-outcome associations).

This chapter describes the exemplar and simulations carried out to examine bias and efficiency when there is missing data in a continuous outcome, focussing in particular on the situation in which this outcome is MNAR. As described in Chapters 1 and 2, a complete case analysis will give biased estimates of the exposure-outcome association under this condition (i.e. if an outcome variable is MNAR conditional on the covariates included in the analysis model). In addition, IPW, FIML, and a standard implementation of MI will also produce biased estimates of the exposure-outcome association.

The work presented in this chapter has been previously published: Cornish RP et al. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. Int J Epidemiol 2015; 44(3):937-45 https://doi.org/10.1093/ije.dyv035 and Cornish RP et al. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. Emerg Themes Epidemiol 2017 14:14. https://doi.org/10.1186/s12982-017-0068-0.

5.1 Exemplar: duration of breastfeeding and IQ

This exemplar is described in Chapter 2 (Section 2.5.1) and investigates whether increased duration of breastfeeding is associated with higher IQ.

5.1.1 Analysis

Subjects included in this analysis were all singletons and twins whose mothers enrolled into ALSPAC in the initial recruitment phase and who were alive at one year (n=13,975).

5.1.1.1 Variables

The variables used in this analysis are listed in Table 5-1 below (they are described in detail in Chapter 3).

Table 5-1: Variables included in the analysis of Exemplar 1

Variable	Details	Chapter 3 section[1]
Outcome	IQ at 15 years	3.3.1
Exposure	Duration of breastfeeding: this was categorised, with those who reported breastfeeding for less than one month combined with those who never breastfed.	3.3.2
Covariates	Child's sex; maternal age (at the child's birth and at first pregnancy), parity, marital status, depression score during pregnancy, ethnicity and smoking; mother's and father's educational level; family occupational social class; housing tenure; use of car; phone in home; double glazing; number of rooms in the house and financial difficulties score.	3.3.3
Linked education variables	Attainment: KS4 capped point score (range 0-540) Number of A*- C grades at GCSE/equivalent, dichotomised as <5 or 5 or more. KS3 attainment score (range 0-430) KS2 attainment score (range 0-280) Other: Percent absence in Year 11 (last year of KS4) SEN status in Year 11	3.4.1

1. Summary statistics for each variable are given in Section 3.5

The analysis model is a linear regression of IQ on duration of breastfeeding and the covariates listed in Table 5-1. The linked education variables are used as auxiliary variables in the MI and FIML models, as described below.

5.1.1.2 Association between IQ and attainment

The association between IQ and attainment was first investigated (among all those with complete data on IQ and the relevant attainment variable) by examining scatterplots. Since these relationships appeared from the graphs to be non-linear, fractional polynomials were used to select the most appropriate model (Royston et al. 1999). The best fitting one-power fractional polynomial predicted IQ from: KS4 attainment score cubed, KS3 attainment score squared and KS2 attainment score squared. For the KS4 and KS2 scores, the best fitting two-power model did not have a significantly better fit than the one-power model (KS4: change in deviance = 34380.1-34376.1=4.0, p=0.1; KS2: change in deviance = 32321.6 – 32316.9=4.7, p=0.1). Therefore, the one-power models were chosen. For the KS3 attainment score, the best fitting two-power model (powers 3 and -0.5) did fit significantly better than the one power model (change in deviance = 28087.6-28072.0=15.5, p<0.001). However, for simplicity (in terms of carrying out the imputations and inclusion in the FIML models) the best fitting one power model was also used for this variable.

Further, because I hypothesised that the relationship between attainment and IQ might differ by socio-economic position, I also examined the relationship separately for mothers with low educational attainment (O level or lower) and those with higher attainment (A level or degree/higher).

5.1.1.3 Examining the missing data mechanism

Logistic regression was used to examine the predictors of missingness in IQ. The factors included were those identified in Chapter 4; the linked education variables were included to investigate whether there was evidence for IQ being MNAR.

5.1.1.4 Dealing with missing data

Four different approaches were used to deal with missing data when modelling the relationship between breastfeeding and IQ:

a) A complete case analysis
b) Inverse probability weighting, both including and excluding the linked education variables.
c) Multiple imputation using chained equations, also performed both with and without the linked education variables.
d) Full information maximum likelihood, incorporating the linked variables.

5.1.1.4.1 Inverse probability weighting

This method is described in Section 2.1.2. I used two (logistic) models to obtain the inverse probability weights. Model 1 included all the baseline covariates (including the exposure, duration of breastfeeding); model 2 also included all linked education variables. The IPW models using auxiliary variables included only individuals with fully observed covariates and auxiliary variables. The Hosmer-Lemeshow goodness of fit test (Hosmer 1989) was used to assess the fit of the logistic models used to generate the (inverse probability) weights. As a sensitivity analysis, large weights were truncated – choosing different maximum values (8, 6 and 4).

5.1.1.4.2 MI models

Multiple imputation is described in detail in Section 2.1.3. I used two different imputation models to carry out MI; both models included all variables in the analysis model. In addition to these, the second set of imputations included the linked education variables as auxiliaries. For each analysis, 100 datasets were imputed with 10 burn-in iterations (cycles). In the MI models incorporating linked education data, IQ was imputed from the KS4 attainment score cubed, the KS3 attainment score squared and the KS2 attainment score squared. Similarly, the KS4 attainment score was imputed from the cube root of IQ and the KS3 and KS2 attainment scores from the square root of IQ. The equations for all other variables included each attainment

score as a linear variable. The dichotomous KS4 attainment variable, SEN status and school absence (squared rooted) were also included in these MI models.

The main reason for individuals having missing education data is because they attended an independent school (as described in Section 3.2.1). This is because contribution to the NPD is only compulsory for schools that follow the national curriculum, so individuals who attended an independent school at the time of the linkage would not have been linked unless their school contributed data voluntarily. Therefore, as a sensitivity analysis I deducted ten IQ points from the imputed IQs of all individuals with missing attainment data (i.e. individuals with missing IQ and attainment). This ad hoc method for carrying out a sensitivity analysis was suggested by Rubin (Rubin 1987). I did this under the assumption that their imputed IQ would be an over-estimate because children who attend independent schools would, on average, obtain better GCSE grades for a given IQ.

5.1.1.4.3 FIML

The FIML was carried out with the KS4, KS3 and KS2 attainment scores and the absence variable as extra dependent variables (Graham 2003). FIML, including how auxiliary variables are incorporated in FIML, is outlined in Section 2.1.4. Each attainment score was squared and standardised for this analysis; similarly, the square root of the absence score was also standardised. IQ was left as raw scores. These transformations were done to meet the assumption of multivariate normality (Kline 2011). Neither the dichotomous attainment variable nor the SEN variable was included in this analysis. Thus, the regression models specified were:

$$\text{IQ} = \mathbf{X}\beta + \epsilon_{IQ}$$

$$\text{KS4}_z^2 = \mathbf{X}\,\partial + \epsilon_{KS4_z^2}$$

$$\text{KS3}_z^2 = \mathbf{X}\gamma + \epsilon_{KS3_z^2}$$

$$\text{KS2}_z^2 = \mathbf{X}\omega + \epsilon_{KS2_z^2}$$

$$\text{sqrt(ABS)}_Z = \boldsymbol{X}\tau + \epsilon_{sqrt(ABS)_z}$$

where X refers to the covariates included in the analysis model, and β refers to the vector of regression coefficients from the model of IQ against these covariates, and δ, γ, ω, and τ to the vectors of regression coefficients from the models of KS4 attainment, KS3 attainment, KS2 attainment, and square root of percent absence against these same covariates (respectively). The errors ϵ_{IQ}, $\epsilon_{KS4_z^2}$, $\epsilon_{KS3_z^2}$, $\epsilon_{KS2_z^2}$ and $\epsilon_{sqrt(ABS)_z}$ were specified as being correlated, as outlined in Chapter 2 (Section 2.1.4.1).

5.1.2 Results

Of the 13,975 subjects included in this analysis (singletons and twins enrolled in the initial recruitment phase who were alive at one year), 12,565 had breastfeeding data, 12,038 had at least one linked education variable (9,100 had complete linked data), 4,918 had non-missing breastfeeding and IQ data, and 3,696 were complete cases (individuals with breastfeeding, IQ and complete covariate information, but not necessarily linked data). Table 5-2 gives the numbers with and without linked data according to completeness of ALSPAC data.

Table 5-2: Completeness of ALSPAC data by availability of linked data

Complete data on:			Linked data		Total
Covariates	Breastfeeding	IQ	Yes[1]	No	
Yes	Yes	Yes	3,327	359	3,686
		No	3,883	670	4,553
	No	Yes	26	4	30
		No	181	44	225
No	Yes	Yes	1,118	114	1,232
		No	2,595	499	3,094
	No	Yes	64	11	75
		No	844	236	1,080
			12,038	1,937	13,975

1. At least one linked variable (of the 12,038 with at least one linked variable, 9,100 (76%) had complete linked education data).

Note that age at first pregnancy was included in an initial analysis. However, since it was strongly associated with maternal age, resulting in the MI models not converging, and because I found no evidence that it was associated with either IQ (difference

between lowest and highest age at first pregnancy = 0.8 IQ points; p= 0.8 for overall association) or missingness in IQ (ORs between 0.99 and 1.03), I omitted this variable from the analysis. Thus, the numbers in Table 5-2 above refer to those with all covariates not including age at first pregnancy.

5.1.2.1 Association between IQ and attainment

Graphs of IQ against the attainment scores are shown in Figure 5-1. As described above (Section 5.1.1.2), the best fitting one power model predicted IQ from the KS4 attainment score cubed, KS3 attainment score squared and KS2 attainment score squared. Table 5-3 shows the bivariate correlations between these (transformed) attainment scores and IQ. Among the 3,636 individuals with IQ and all three attainment scores, the three attainment scores together explained just under 44% of the variability in IQ (R^2 = 0.437).

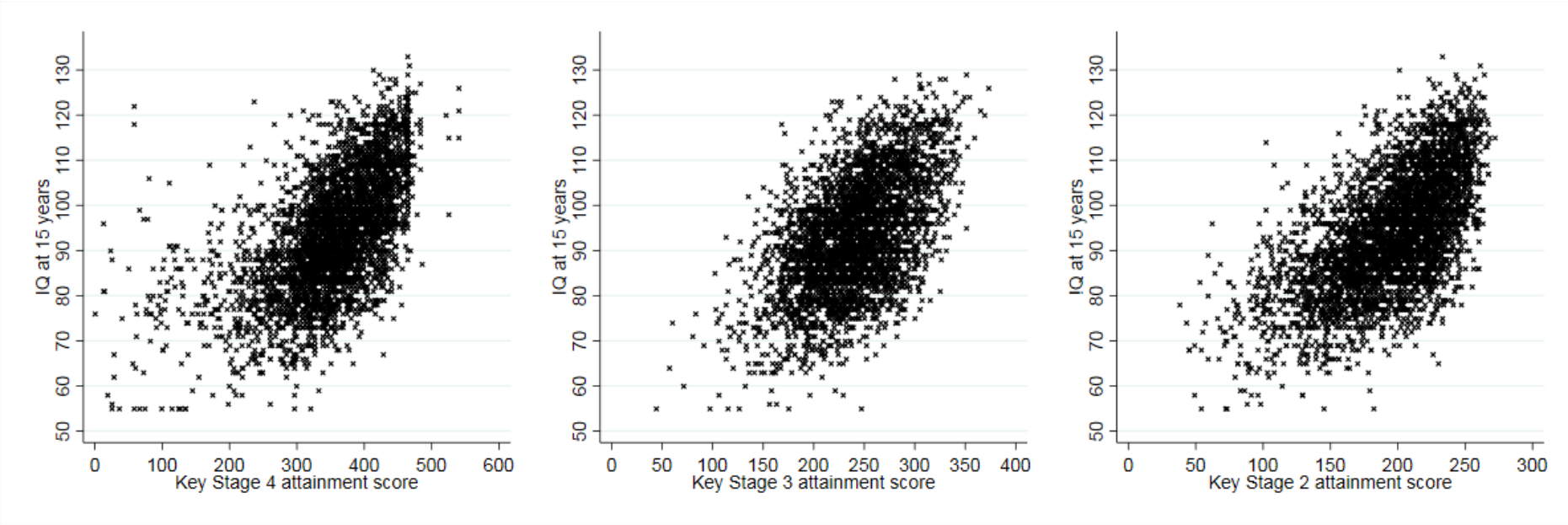


Figure 5-1: Plot of IQ against the KS4, KS3 and KS2 attainment scores

Table 5-3: Correlations between transformed attainment scores and IQ[1]

		Attainment score	
	IQ	KS4 cubed	KS3 squared
KS4 cubed	0.60	---	---
KS3 squared	0.53	0.68	---
KS2 squared	0.62	0.74	0.64

1. Among the 3,636 individuals with data for IQ and all attainment scores

5.1.2.2 Predictors of missing IQ

Altogether 5,023 of the 13,975 individuals (36%) had IQ data. Tables 5-4 to 5-6 show predictors of missingness in IQ among individuals with breastfeeding, complete covariate information and all linked data (n=5,554). Note that the main driver of this reduction from 13,975 to 5,554 was availability of the linked education data (although 12,038 individuals had at least one linked education variable, only 9,100 had all five). Females were less likely to have missing IQ, as were first born children and those born to mothers who did not smoke, were older, and had higher socio-economic position (according to education and several other factors). Increased duration of breastfeeding was also associated with lower odds of having missing IQ data. After mutually adjusting for all other factors, there was little evidence that father's education (OR=1.01 and 1.06 for A levels and degree/higher compared to O levels/lower) and maternal depression (OR=1.00) were predictive of missing IQ. Lower attainment (particularly the Key Stage 4 capped point score) and higher school absence were strongly related to having missing IQ data, even after adjusting for the baseline covariates (Table 5-6).

Table 5-4: Predictors of missingness in IQ: child and maternal covariates (n=5,554 with complete covariate information plus complete linked data)

Factor	Level	OR (95% CI)[1]	p-value
Sex	Female vs male	0.83 (0.74, 0.94)	p=0.003
Mother's education	O level or lower	1.00	
	A level	0.81 (0.70, 0.93)	
	Degree or higher	0.84 (0.67, 1.05)	p=0.009
Duration of breastfeeding	Never / < 1 month	1.00	
	1 to <3 months	0.76 (0.64, 0.91)	
	3 to <6 months	0.71 (0.60, 0.83)	
	6 months +	0.59 (0.51, 0.68)	p<0.001
Mother's age at birth	<20	1.00	
	20-24	0.85 (0.53, 1.36)	
	25-29	0.73 (0.45, 1.17)	
	30-34	0.61 (0.38, 0.99)	
	35+	0.53 (0.32, 0.89)	p=0.002
Smoking	Never	1.00	
	Yes; not in pregnancy	1.27 (1.11, 1.45)	
	Yes; in pregnancy	1.48 (1.26, 1.74)	p<0.001
Parity	0	1.00	
	1	1.31 (1.15, 1.50)	
	2	1.63 (1.33, 1.97)	
	3+	1.87 (1.37, 2.57)	p<0.001
Mother's marital status	Not married vs married	0.89 (0.75, 1.06)	p=0.2
Mother's ethnicity	Non-white vs white	0.71 (0.44, 1.15)	p=0.2
Maternal depression (EPDS)	Per 1 point increase	1.00 (0.99, 1.02)	p=0.6

1. Mutually adjusted for all covariates plus all linked education variables

Table 5-5: Predictors of missingness in IQ: family/paternal covariates
(n=5,554 with complete covariate information plus complete linked data)

Factor	Level	OR (95% CI)[1]	p-value
Father's education	O level or lower	1.00	
	A level	1.01 (0.88, 1.15)	
	Degree or higher	1.06 (0.87, 1.29)	p=0.8
Housing tenure	Mortgaged /owned	1.00	
	Private rented	1.31 (0.97, 1.77)	
	Council/HA/other	0.97 (0.77, 1.20)	p=0.2
Number of rooms	Per 1 room increase	0.94 (0.89, 0.99)	p=0.02
Phone in home	No vs yes	1.30 (0.97, 1.74)	p=0.08
Car use	No vs yes	1.16 (0.85, 1.58)	p=0.3
Double glazing	No vs full/partial	1.03 (0.92, 1.16)	p=0.6
Financial difficulties	Per 1 unit increase	(1.00, 1.03)	p=0.1
Family occupational social class	IIIm-V vs I-IIIn	1.14 (0.96, 1.35)	p=0.1

1. Mutually adjusted for all covariates plus all linked education variables

Table 5-6 Predictors of missingness in IQ: attainment variables
(n=5,554 with complete covariate information plus complete linked data)

Factor	Level	OR (95% CI)[1]	p-value
Five or more A-C grades at Key Stage 4	No	1.00	
	Yes	0.95 (0.79, 1.15)	p=0.6
Key Stage 4 score	OR for each 10 point increase	0.97 (0.95, 0.98)	p<0.001
Key Stage 3 score		0.99 (0.97, 1.00)	p=0.1
Key Stage 2 score		0.99 (0.96, 1.01)	p=0.2
Square root % absence	Per 1 unit increase	1.12 (1.07, 1.18)	p<0.001
SEN status	None	1.00	
	School action	1.06 (0.86, 1.31)	
	Statement	1.28 (0.61, 2.66)	p=0.7

1. Mutually adjusted for all covariates plus all linked education variables

5.1.2.3 Predictors of missing linked education data

In Chapter 4 I showed that individuals whose parents were more highly educated were less likely to have linked education data, as were males, those breastfed for longer, and first-born children. Missingness in the linked education data was also associated with other indicators of socio-economic position, with those from more advantaged backgrounds being more likely to have missing education data. In the present analysis I found that, after adjustment for all these baseline covariates considered in Chapter 4, missingness in linked education data was associated with higher IQ (OR for each 1 point increase = 1.01, 95% CI 1.00, 1.02, p=0.006).

5.1.2.4 Relationship between duration of breastfeeding and IQ

Table 5-7 gives the predicted mean difference in IQ score for increasing duration of breastfeeding obtained using the different approaches. In the complete case analysis, the (fully adjusted) estimated mean difference in IQ for the three breastfeeding groups (compared to never / <1 month), were 0.8, 2.6 and 3.5 (confidence intervals shown in the table). In the IPW and MI analyses when the linked education variables were not included, estimates for 1 to <3 months and 3 to <6 months were quite similar to the complete case estimates; there were greater differences for 6+ months (3.7 in the IPW analysis and 3.2 in the MI analysis). The MI estimates had slightly narrower confidence intervals than the complete case estimates whereas the IPW estimates were slightly less precisely estimated (compared to the complete case estimates).

In the analyses that included the linked education data, all three approaches (IPW, MI and FIML) gave broadly similar results in terms of point estimates, although the adjusted odds ratios from IPW were slightly higher than those obtained using MI or FIML. In all cases, these estimates were higher than those obtained using the complete case analysis and the other analyses that did not include the linked education data. Using IPW resulted in a loss of precision relative to the complete case analysis; this was largely due to fact that this analysis only included complete cases

for whom all linked education variables were available. In contrast, use of MI and FIML resulted in gains in efficiency.

The Hosmer-Lemeshow tests did not indicate a poor fit in the (IPW) models used to predict missing IQ (χ_8^2=8.1, p=0.4 when linked education data were excluded and χ_8^2=3.6, p=0.9 when they were included in the model).

Table 5-7: Relationship between duration of breastfeeding and IQ: estimates obtained from different analysis approaches

		Analysis approach					
			Excluding linked data		Including linked data		
	Duration of Breastfeeding	Complete case analysis (n=3,686)	Inverse probability weighting (n=3,686)	Multiple imputation (n=13,975)	Inverse probability weighting (n=2,643)	Multiple imputation[2] (n=13,975)	FIML[3] (n=13,693/ 13,975)[4]
Unadjusted results	Never / < 1 month	---	---	---	---	---	---
	1 to <3 months	2.0 (0.6, 3.4)	1.9 (0.2, 3.5)	2.7 (1.6, 3.8)	2.6 (0.9, 4.3)	3.5 (2.5, 4.5)	3.5 (2.5, 4.6)
	3 to <6 months	5.1 (3.8, 6.3)	5.5 (4.2, 6.7)	6.0 (4.9, 7.1)	5.0 (3.5, 6.5)	6.9 (5.9, 7.8)	6.9 (6.0, 7.9)
	6 months +	7.6 (6.6, 8.6)	8.1 (7.0, 9.1)	8.4 (7.5, 9.3)	8.2 (7.0, 9.5)	9.6 (8.9, 10.3)	9.6 (8.9, 10.4)
Adjusted results[1]	Never / < 1 month	---	---	---	---	---	---
	1 to <3 months	0.8 (-0.5, 2.1)	0.7 (-0.8, 2.2)	0.8 (-0.3, 1.9)	1.6 (0.0, 3.1)	1.3 (0.4, 2.3)	1.4 (0.4, 2.3)
	3 to <6 months	2.6 (1.4, 3.8)	2.7 (1.5, 3.9)	2.5 (1.4, 3.5)	3.0 (1.7, 4.4)	3.0 (2.0, 3.9)	3.0 (2.1, 3.9)
	6 months +	3.5 (2.5, 4.5)	3.7 (2.6, 4.8)	3.2 (2.3, 4.1)	4.4 (3.2, 5.7)	3.9 (3.1, 4.7)	3.9 (3.1, 4.7)
For adjusted results:							
Gain in precision[5]	Never / < 1 month	---	---	---	---	---	---
	1 to <3 months	N/A	-26%	38%	-34%	82%	72%
	3 to <6 months		-7%	20%	-25%	58%	65%
	6 months +		-14%	19%	-36%	66%	63%

1. Adjusted for sex, maternal and paternal education, social class, parity, maternal depression, age, marital status, ethnicity & smoking, housing tenure, number of rooms in house, whether house had double glazing, use of car and phone, and financial difficulties score
2. IQ predicted from KS4 points cubed, KS3 points squared and KS2 points squared (best fitting fractional polynomials of degree 1), plus all other factors.
3. Z-scores of square root of percent absence, and KS2, KS3 and KS4 points squared were added as extra dependent variables
4. The crude analysis included 13,693 individuals (excluded 282 individuals with breastfeeding, IQ, absence, KS2, KS3 and KS4 data missing); the adjusted analysis included all 13,975 individuals
5. Relative to complete case analysis

5.1.2.4.1 Sensitivity analyses

IPW analysis

When large weights were truncated in IPW, the estimated effect of breastfeeding was slightly reduced as the maximum value of the weights was decreased (i.e. from 8 to 4). For example, when truncating the weights at 6, the fully adjusted estimates (using the linked variables) were 1.3, 2.9 and 4.2. (More detailed results are given in Appendix B, Table 2). There was also an increase in precision – the variances were between 4% and 26% larger in the (fully adjusted) models without weights truncated compared to the model when weights were truncated to a maximum value of 4.

MI analysis

Deducting IQ points from individuals with imputed GCSE scores made very little difference to the results (Table 5-8).

Table 5-8: Sensitivity analysis: deducting 10 points from imputed IQs when linked data was missing

Breastfeeding duration	Unadjusted	Adjusted[1]
Never / < 1 month	--	--
1 to <3 months	3.5 (2.5, 4.5)	1.3 (0.3 2.3)
3 to <6 months	6.8 (5.8, 7.8)	2.9 (1.9, 3.8)
6 months +	9.5 (8.8, 10.3)	4.0 (3.2, 4.7)

1. Adjusted for sex, maternal and paternal education, social class, parity, maternal depression, age, marital status, ethnicity & smoking, housing tenure, number of rooms in house, whether house had double glazing, use of car and phone, and financial difficulties score

5.1.3 Discussion

As discussed in Chapters 1 and 2, in a linear regression, if missingness depends on the outcome (given the covariates) then both a complete case analysis and a standard implementation of MI will produce biased estimates of the exposure-outcome association (Carpenter and Kenward 2013, White and Carlin 2010), as will an analysis using FIML or IPW.

In the exemplar described above I showed that missingness in IQ (the outcome) was associated with duration of breastfeeding (the exposure). I then showed that school

attainment (used as proxies for IQ) was also associated with missingness in IQ, suggesting that IQ was indeed MNAR. Inclusion of the linked attainment variables – strong predictors of both IQ and missingness in IQ – in the IPW, FIML and MI models would mean that IQ would be more likely to be MAR, although this assumption is not testable. In this situation, FIML, IPW and MI would give unbiased estimates of the breastfeeding-IQ association. However, because of the relatively large amount of variation in IQ for a given level of attainment, it is possible that IQ remained MNAR, albeit to a lesser extent, even after including attainment; as such, some bias could remain. It is not possible to determine the extent of this. The IPW estimates (including the linked education variables) were slightly higher than the FIML and MI estimates, although were reduced after truncating large weights. Again, it is not possible to determine from the data which results are likely to be the least biased.

5.2 Simulations

This simulation study was based on the above exemplar and was carried out to investigate the impact of missingness in a continuous outcome on the exposure-outcome relationship – particularly in terms of bias, but also on efficiency – and to quantify the extent to which this bias can be reduced with the inclusion of linked variables. The key factors that were varied were the degree of correlation between the original outcome (IQ) and its linked proxy (attainment), the proportion of missing data and the extent to which the outcome was MNAR. Complete data were simulated first; missing data were then simulated in a separate process.

5.2.1 Simulated datasets

The following four variables were all simulated, with distributions chosen to be roughly representative of those seen in ALSPAC: offspring sex and mother's education, the exposure variable (duration of breastfeeding) and the outcome variable (IQ). For simplicity I simulated a single proxy (attainment) variable; I thought of this as analogous to the linked KS4 attainment score, as this was the continuous attainment variable measured at a similar age to IQ. I will refer to this as the linked

attainment score in the remainder of this chapter. I simulated datasets of 10,000 observations – to approximately match the numbers in ALSPAC with complete baseline covariates. Sex and mother's education were the two covariates. Sex was drawn from a Bernoulli distribution with probability 0.5 and mother's education from a multinomial random variable with probabilities 0.5, 0.25 and 0.25, corresponding to categories O level or lower, A level, and degree or higher (respectively). The exposure variable, duration of breastfeeding, was created as a categorical variable, with categories designed to represent: never/less than one month, 1-<3 months, 3-<6 months, and 6+ months. This variable was simulated as a series of multinomial distributions, conditional on mother's education such that duration of breastfeeding increased with higher maternal education. The marginal probabilities for the four breastfeeding categories were: (0.5,0.15,0.15,0.2), (0.3,0.1,0.2,0.4) and (0.15,0.1,0.15,0.6) for O level/lower, A level, and degree/higher, respectively. The outcome, IQ at age 15 years, was simulated as a standard normal variable, dependent on sex, mother's education and duration of breastfeeding such that:

$$\begin{aligned} \text{IQ}_i = \ & \beta_0 + \beta_1 \times sex_i + \beta_2 \times mumed_{1i} + \beta_3 \times mumed_{2i} \\ & + \beta_4 \times BF_{1i} + \beta_5 \times BF_{2i} + \beta_6 \times BF_{3i} + \varepsilon_i \end{aligned} \qquad [1]$$

This was also the analysis model. In this equation, sex is the indicator variable for sex, mumed_1 and mumed_2 for the mother having A levels and a degree level qualification or higher, respectively, BF_1, BF_2 and BF_3 are the indicator variables for being breastfed for 1 to <3, 3 to <6 and 6 months or longer, and ε is the random error, following a normal distribution with mean 0 and variance σ^2, with the latter calculated to give IQ a variance of 1. In this equation and throughout the remainder of the thesis, the subscript i denotes an individual. The coefficients of this regression model were fixed to be as follows: $\beta_0 = -0.4$, $\beta_1 = -0.1$, $\beta_2 = 0.4$, $\beta_3 = 0.8$, $\beta_4 = 0.1$, $\beta_5 = 0.2$, $\beta_6 = 0.3$, representing relationships similar to those seen in ALSPAC. The linked attainment score was also simulated as a standard normal variable. For simplicity, this was made dependent (with a linear relationship) only on IQ:

$$KS4_i = \rho \times IQ_i + \tau_i \quad [2]$$

with KS4 representing the Key Stage 4 attainment score and τ the (normal) random error with mean 0 and variance φ^2, again calculated to give the attainment score a variance of 1. In a sensitivity analysis, the attainment score was made dependent on sex and mother's education in addition to IQ.

5.2.2 Simulating the missing data

Because the focus of this investigation was the utility of proxy data for missing outcomes, I only created missing data in the outcome variable, IQ. The probabilities were initially generated using a logistic regression model, with coefficients similar to those seen in ALSAPC. However, this led to some unexpected results – in which the percent bias was highest (and similar) when there was 40% and 60% missing data and lowest (and again similar) when there was 20% and 80% missing data. A limited set of complete case and MI results (estimates, standard errors and % bias) are given in Appendix B, Table 3. I investigated this further by simulating single datasets with 20%, 40%, 60% and 80% missing data and found that the change in pseudo R^2 when adding the outcome (IQ) to a missingness model (logistic regression predicting the log odds of being observed) containing sex, mother's education and duration of breastfeeding was similar when there was 20% and 80% missing data (change in pseudo R^2 = 5.2% in both cases) and lower than when there was 40% and 60% missing data (change in pseudo R^2 = 6.2% for 40% missing data and 6.3% for 60% missing data). Because of this apparent symmetry, I decided to generate the probabilities (of being observed) using a binomial regression model instead of a logistic model. The coefficients were again similar to those seen in ALSPAC:

$$\begin{aligned} \mathrm{P(observed)}_i = {} & \alpha + \gamma_1 \times sex_i + \gamma_2 \times mumed_{1i} + \gamma_3 \times mumed_{2i} \quad [3] \\ & + \gamma_4 \times BF_{1i} + \gamma_5 \times BF_{2i} + \gamma_6 \times BF_{3i} + \gamma_7 \times IQ_i \\ & + \gamma_8 \times BF_{1i} \times IQ_i + \gamma_9 \times BF_{2i} \times IQ_i + \gamma_{10} \times BF_{3i} \times IQ_i \end{aligned}$$

Some of the regression coefficients in this model were fixed throughout the simulation study: γ_1 = 0.04 (female compared to male), γ_2=0.075 (mother's education = A level compared to mother's education = O level or lower), γ_3=0.10 (mother's education degree or higher compared to O level or lower), γ_4=0.08 (breastfed for 1 to <3 months compared to never/less than one month), γ_5=0.12 (breastfed for 3 to <6 months compared to never/<one month) and γ_6=0.14 (breastfed for 6+ months compared to never/<one month). The remaining coefficients were varied. The values of α were first calculated exactly and then adjusted, where necessary, using a trial and improvement method in order to produce particular percentages of missing data (20%, 40%, 60% or 80%). These adjustments using trial and improvement were necessary for scenarios in which Equation 3 produced negative predicted probabilities or predicted probabilities that were greater than one.

For each observation a Bernoulli random variable with p=Pr(IQ observed), as given by Equation 3, was drawn to determine whether each IQ was missing. However, because a binomial model was used to predict the probability of IQ being observed, this sometimes led to negative predictions or predictions that were greater than one; this was particularly the case when simulating datasets with 80% missing data. When the probability was predicted as negative or greater than one, IQ was automatically set to missing or observed (respectively).

Finally, I simulated the linked attainment score to be missing (specifically MNAR), with missingness only dependent on itself; again, the probabilities were generated using a binomial regression model:

$$\text{P(link observed)}_i = \pi + \delta \times KS4_i \qquad [4]$$

The value of the intercept, π, was set at 0.8 in order to generate 20% missing attainment data. Values of δ were varied.

5.2.2.1 Scenarios

Key factors influencing the extent of bias are the amount of missing data and the degree to which the outcome is MNAR; the strength of association between the outcome and its proxy will largely determine the degree to which this bias can be reduced. Thus, these constituted the three primary factors varied in the simulations. These were varied as detailed below.

Factor 1: The percentage of missing outcome (IQ) data: 20%, 40%, 60%, 80%.
Factor 2: How good a proxy the linked variable was: correlation between the outcome variable (IQ) and its linked proxy (attainment score) = 0.1, 0.3, 0.5, 0.7, 0.9.
Factor 3: Whether the outcome (IQ) was MAR or MNAR and, if the latter, the extent of this: increase in probability of observing IQ for a one SD increase in IQ = 0 (MAR), 0.05, 0.1, and 0.2 (γ_7 from Equation [3]).

In addition, I hypothesised that if the association between IQ and the probability of it being missing varied according to duration of breastfeeding, this would substantially increase bias. Thus, the first secondary factor was:

Factor 4: Whether or not the association between the outcome (IQ) and the probability of it being observed differed according to the exposure (breastfeeding): γ_8 from Equation [3] = 0 or -0.025. For simplicity, I made the strength of association between IQ and the probability of it being missing change linearly with increasing breastfeeding, such that:

$$\gamma_9 = 2\gamma_8 \text{ and } \gamma_{10}=3\gamma_8 \text{ (from Equation [3])}$$

Finally, in the main sets of scenarios there was no missingness in the linked variable. However, I also wanted to consider some scenarios in which the linked proxy was not available for all individuals and therefore introduced missingness in this variable. This formed the other secondary factor:
Factor 5: Whether or not there was missingness in the linked attainment score and varying the direction of missingness: difference in probability of Pr(KS4 observed) for a one SD increase in attainment score (KS4) = -0.10, +0.10.

The scenarios are summarised in Table 5-9. I did not consider every possible combination of these factors. The main set of scenarios involved only the three primary factors listed above. However, at each of the four levels of missing data the MAR condition was only simulated in one scenario, with a correlation of 0.7 between IQ and the linked attainment score. This was because the focus in this study was primarily on reducing bias with an outcome variable that is MNAR. I included MAR in the simulations simply to show that the complete case analysis and MI would both be unbiased and that, if auxiliary variables were included, MI would simply increase precision in this situation. Additional scenarios involved the two secondary factors but these were only introduced for a limited set of scenarios. Altogether, there were 100 scenarios. For each scenario, 1000 datasets were simulated.

Table 5-9: Scenarios investigated in the simulations based on Exemplar 1

			Factor 3: Change in Pr(IQ observed) for one SD increase in IQ	Factor 2: Correlation between IQ and linked attainment score (KS4)				
				0.1	0.3	0.5	0.7	0.9
Main set of scenarios (each at 20%, 40%, 60%, 80% missing IQ (Factor 1)): 64 scenarios								
			0				✓	
			0.05	✓	✓	✓	✓	✓
			0.1	✓	✓	✓	✓	✓
			0.2	✓	✓	✓	✓	✓
Secondary sets of scenarios (each at 20%, 40%, 60%, 80% missing IQ (Factor 1)): 24 scenarios								
Missing linked data	Factor 5: Change in Pr(KS4 observed) for one SD increase in KS4	Factor 4: Association between IQ and Pr(IQ observed) dependent on breastfeeding?						
No	---	Yes	0.1[1]	✓	✓	✓	✓	✓
Yes, 20%	-0.05	No	0.1				✓	
	-0.10		0.1				✓	
	+0.05		0.1				✓	
	+0.10		0.1				✓	

1. In the baseline breastfeeding group; reduction in this coefficient of 0.025 for each consecutive breastfeeding group

5.2.3 Statistical Analysis

As in the analysis of the ALSPAC data described above, I estimated the coefficients for breastfeeding (β_4, β_5 and β_6) using the multiple linear regression model given by Equation [1]:

$$\begin{aligned} \text{IQ}_i = \ & \beta_0 + \beta_1 \times sex_i + \beta_2 \times mumed_{1i} + \beta_3 \times mumed_{2i} \\ & + \beta_4 \times BF_{1i} + \beta_5 \times BF_{2i} + \beta_6 \times BF_{3i} + \varepsilon_i \end{aligned} \qquad [1]$$

These (β_4, β_5, and β_6) were estimated using:

a) A complete case analysis
b) Inverse probability weighting. The logistic models used to predict the probability of missingness included breastfeeding, the covariates (sex and mother's education) plus the linked attainment score. When an interaction (between breastfeeding and IQ) was introduced in the missingness model, these models also included an interaction between breastfeeding and the linked attainment score. However, I also carried out a second set of IPW analyses in which this interaction term was omitted from the models used to generate the weights.
c) Multiple imputation. For each simulated dataset, 100 imputed datasets were created. The imputation models included all the variables in the analysis model plus the linked attainment score.
d) Full information maximum likelihood via structural equation modelling. This was done by specifying the linked attainment score as a second dependent variable. Thus, the regression models specified were:

$$\begin{aligned} \text{IQ}_i = \ & \beta_0 + \beta_1 \times sex_i + \beta_2 \times mumed_{1i} + \beta_3 \times mumed_{2i} \\ & + \beta_4 \times BF_{1i} + \beta_5 \times BF_{2i} + \beta_6 \times BF_{3i} + \varepsilon_{IQ\_i} \end{aligned}$$

and

$$\begin{aligned} \text{KS4}_i = \ & \delta_0 + \delta_1 \times sex_i + \delta_2 \times mumed_{1i} + \delta_3 \times mumed_{2i} \\ & + \delta_4 \times BF_{1i} + \delta_5 \times BF_{1i} + \delta_6 \times BF_{1i} + \varepsilon_{KS4\_\text{i}} \end{aligned}$$

with the errors ε_{IQ} and ε_{KS4} specified as being correlated.

The estimates obtained from these analyses were compared to the true parameters 0.1, 0.2 and 0.3. For each parameter, β_j, the bias was estimated as $\overline{b_j} - \beta_j$, where $\overline{b_j}$ is the estimated regression coefficient for parameter β_j averaged over the 1000 simulated datasets. This was converted to percentage bias. I also calculated the empirical standard error, the standard deviation of the point estimates for each parameter. In addition, for the analyses using multiple imputation, I calculated the fraction of missing information (FMI) for each coefficient (Rubin 1987) and the percent increase in precision compared to the complete case analysis; the latter is given by the variance of the point estimates for the parameter of interest obtained using a complete case analysis divided by the variance obtained using multiple imputation. The percent increase in precision (relative to the complete case analysis) is also given for the IPW and FIML results.

5.2.4 Results

5.2.4.1 Outcome MAR

As expected, when the data were simulated as MAR (with the variables related to missingness being included in the analysis model) all three analyses gave unbiased estimates. Multiple imputation – with the linked attainment score as an auxiliary variable – increased precision in all cases where the data were MAR, although the increases were relatively small when the percentage of missing information was low (Table 5-10). Similarly, FIML increased precision when there was 40%, 60% and 80% missing data but not when there was only 20% missing (Table 5-10). IPW generally resulted in slight losses in precision, particularly with 80% missing data.

Table 5-10: Estimates of β_4, β_5 and β_6 when outcome (IQ) simulated as MAR (true values 0.10, 0.20, 0.30, respectively)

		Complete case		MI using linked attainment score (KS4)				IPW with linked attainment score			FIML; attainment score (KS4) as extra dependent variable		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]
IQ 20% missing	β_4	0.1005 (0.033)	0.5%	0.1004 (0.031)	0.3%	10%	15%	0.1005 (0.033)	0.5%	-1%	0.1011 (0.032)	1.1%	2%
Correlation	β_5	0.1990 (0.030)	-0.5%	0.1993 (0.029)	-0.3%	11%	13%	0.1989 (0.030)	-0.5%	-1%	0.1986 (0.030)	-0.7%	-1%
(IQ:KS4) = 0.7	β_6	0.3006 (0.025)	0.2%	0.3002 (0.024)	0.1%	7%	13%	0.3005 (0.025)	0.2%	0%	0.3000 (0.025)	0%	-1%
IQ 40% missing	β_4	0.0994 (0.038)	-0.6%	0.1004 (0.034)	0.3%	22%	30%	0.0995 (0.038)	-0.5%	-1%	0.1014 (0.036)	1.4%	11%
Correlation	β_5	0.1988 (0.035)	-0.6%	0.1996 (0.033)	-0.2%	17%	28%	0.1988 (0.035)	-0.6%	-1%	0.1984 (0.034)	-0.8%	8%
(IQ: KS4) = 0.7	β_6	0.3004 (0.030)	0.1%	0.3004 (0.027)	0.1%	17%	29%	0.3003 (0.030)	0.1%	0%	0.3004 (0.028)	0.1%	14%
IQ 60% missing	β_4	0.1005 (0.049)	0.5%	0.1009 (0.042)	1.1%	34%	50%	0.1002 (0.050)	0.2%	-2%	0.1009 (0.041)	0.9%	41%
Correlation	β_5	0.1975 (0.042)	-1.3%	0.1980 (0.037)	-0.8%	33%	47%	0.1974 (0.042)	-1.3%	-1%	0.1996 (0.038)	-0.2%	19%
(IQ: KS4) = 0.7	β_6	0.2988 (0.037)	-0.4%	0.3002 (0.032)	0.1%	33%	48%	0.2988 (0.037)	-0.4%	1%	0.3002 (0.031)	0.1%	40%
IQ 80% missing	β_4	0.1040 (0.073)	4.0%	0.1050 (0.061)	4.8%	41%	83%	0.1027 (0.075)	2.7%	-7%	0.1024 (0.059)	2.5%	51%
Correlation	β_5	0.2009 (0.062)	0.4%	0.2000 (0.052)	0%	40%	81%	0.1997 (0.063)	0.1%	-6%	0.1999 (0.052)	-0.1%	41%
(IQ: KS4) = 0.7	β_6	0.3011 (0.056)	0.4%	0.3022 (0.046)	0.8%	47%	81%	0.3002 (0.057)	0.1%	-3%	0.3007 (0.045)	0.2%	56%

1. Relative to complete case analysis
2. FMI: Fraction of missing information

5.2.4.2 Outcome MNAR, Factor 1 (% missing data)

When the outcome variable was simulated as MNAR, the complete case, MI, IPW and FIML analyses all produced biased results (as expected); the bias increased as the percentage of missing data increased. Figure 5-2 shows the bias in the first breastfeeding coefficient for the complete case and MI estimates (FIML results were not included on this figure as the results were very similar to those obtained from MI) and further results are shown in Tables 5-11 to 5-14. The comparison between the methods is discussed below in Section 5.2.4.3 as the relative performance of the methods was dependent on the correlation between IQ and its proxy, the attainment score.

The results for 80% missing are likely to have been affected by having negative predictions for the probability of IQ being observed – referred to in Section 5.2.2. In this scenario, an average of 1,156 individuals had a predicted value of Pr(IQ observed) (as given by Equation [3]) that was negative and their IQ was thus set to missing. The IQs among these individuals whose predicted probability of being missing was negative were generally low – the mean of the mean IQs in the 1000 datasets was -1.42 (z-score) and all were below -0.50 – suggesting that, in the scenarios with 80% missing data, the MNAR mechanism was more extreme.

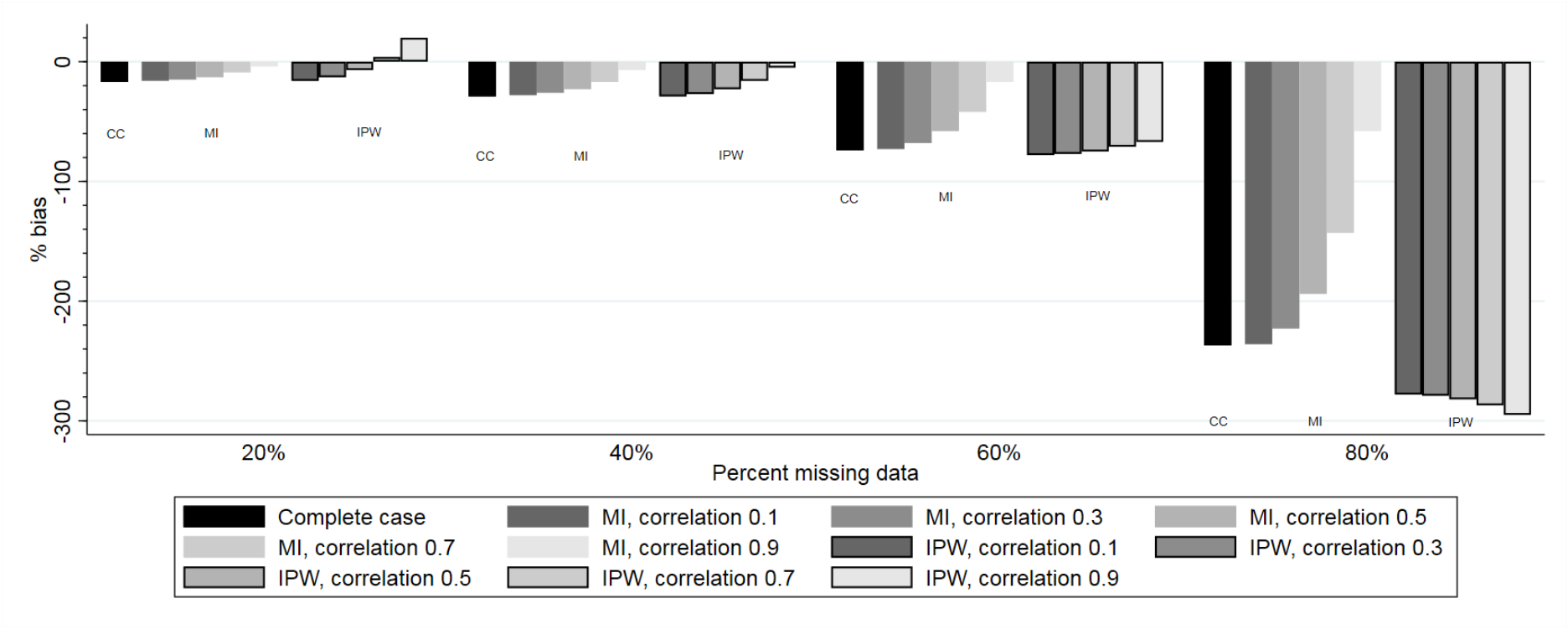


Figure 5-2: Percent bias in first breastfeeding coefficient: complete case, IPW and MI estimates when Pr(IQ observed) = 0.1 for each 1 SD increase in IQ

5.2.4.3 Outcome MNAR, Factor 2 (correlation between outcome and linked proxy)

As the correlation between the outcome and its proxy increased, the amount of information recovered through the imputations increased, thus reducing bias and increasing precision. In summary, in all MNAR scenarios in which the correlation between IQ and the linked attainment score was above 0.1, the results from the MI models were less biased and more precise than the complete case analysis, although the gains were minimal when the correlation was 0.3.

For FIML, the estimates were less biased and more precise for correlations above 0.1 when the percentage of missing data was 40, 60 or 80% but not for 20% missing data; in the latter scenarios the FIML results were only less biased than the complete case analysis when the correlation between IQ and the linked attainment score was at least 0.5. For 20% missing data the FIML results were not generally more precise than the estimates obtained from the complete case analysis.

The performance of IPW (in terms of bias) compared to the other methods varied quite substantially as the percentage of missing data increased. This could be because the missing data model is mis-specified (the simulations used a binomial model whereas I used a logistic model to generate the weights). It could also be partly due to the fact that some of the generated probabilities of being observed were set to zero (because the binomial model used to generate the missing data predicted negative probabilities). With 60% and 80% missing data, the IPW estimates were more biased than those obtained from MI and FIML. They were also more biased than the complete case estimates for all correlations when there was 80% missing data and for correlations of 0.1 and 0.3 when there was 60% missing data. When there was 80% missing data the bias actually increased as the correlation between IQ and attainment increased. For 40% missing data the IPW estimates were very similar to those from both MI and FIML; for 20% missing data the IPW estimates were less biased than the estimates from MI and FIML for correlations of 0.1, 0.3, 0.5 and 0.7 but more biased (and positively rather than negatively biased) when the correlation between IQ and attainment was 0.9.

Changing the correlation between IQ and the linked proxy from 0.5 to 0.9 resulted in reductions of between 12% and 30% in the FMI (Table 5-11 to Table 5-14). These tables also show that the FMI – and the resulting bias – was very similar with 40% missing data and a correlation of 0.7 between the original outcome and its linked proxy as in the scenario with 60% missing data and a correlation of 0.9; similarly, 80% missing data with a correlation of 0.9 resulted in a similar degree of bias to 60% missing data with a correlation of 0.5. With a very good proxy of the original outcome (i.e. with a correlation of 0.9), almost all the bias was eliminated, even with quite high proportions of missing data.

Unsurprisingly, the bias was reduced when the strength of association between IQ and the probability of it being missing was reduced (complete case and MI results only given in Appendix B: Tables 4 and 5) and increased when the strength of this association was increased (complete case and MI results only given in Appendix B: Tables 6 and 7).

Table 5-11: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when IQ MNAR: difference in Pr(IQ observed)=0.10 for 1 SD increase in IQ; 20% missing data

		Complete case		MI				IPW			FIML		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]
IQ 20% missing	β_4	0.08 (0.034)	-17%	0.08 (0.034)	-16%	0%	24%	0.08 (0.035)	-16%	-3%	0.08 (0.035)	-18%	-4%
Correlation	β_5	0.17 (0.030)	-14%	0.17 (0.030)	-14%	1%	21%	0.17 (0.030)	-13%	-1%	0.17 (0.030)	-13%	-1%
(IQ:KS4) = 0.1	β_6	0.26 (0.025)	-12%	0.26 (0.025)	-12%	0%	21%	0.27 (0.025)	-11%	0%	0.26 (0.027)	-13%	-14%
IQ 20% missing	β_4	As above		0.08 (0.034)	-15%	2%	23%	0.09 (0.035)	-13%	-2%	0.08 (0.035)	-17%	-2%
Correlation	β_5			0.18 (0.030)	-13%	4%	20%	0.18 (0.030)	-11%	-1%	0.18 (0.030)	-13%	0%
(IQ: KS4) = 0.3	β_6			0.27 (0.025)	-11%	1%	20%	0.27 (0.025)	-10%	-1%	0.27 (0.027)	-12%	-13%
IQ 20% missing	β_4	As above		0.09 (0.033)	-13%	6%	20%	0.09 (0.035)	-7%	-2%	0.09 (0.034)	-13%	1%
Correlation	β_5			0.18 (0.029)	-11%	8%	18%	0.19 (0.030)	-6%	-2%	0.18 (0.031)	-11%	-5%
(IQ: KS4) = 0.5	β_6			0.27 (0.025)	-9%	4%	18%	0.28 (0.026)	-6%	-1%	0.27 (0.027)	-10%	-14%
IQ 20% missing	β_4	As above		0.09 (0.032)	-9%	14%	16%	0.10 (0.035)	4%	-4%	0.09 (0.033)	-10%	7%
Correlation	β_5			0.19 (0.028)	-7%	14%	13%	0.20 (0.031)	2%	-3%	0.18 (0.030)	-8%	-1%
(IQ: KS4) = 0.7	β_6			0.28 (0.024)	-6%	9%	13%	0.30 (0.026)	0%	-2%	0.28 (0.027)	-7%	-10%
IQ 20% missing	β_4	As above		0.10 (0.031)	-4%	27%	7%	0.12 (0.036)	20%	-9%	0.10 (0.032)	-4%	17%
Correlation	β_5			0.19 (0.027)	-3%	21%	6%	0.23 (0.031)	15%	-7%	0.19 (0.029)	-3%	6%
(IQ: KS4) = 0.9	β_6			0.29 (0.023)	-2%	18%	6%	0.33 (0.026)	10%	-6%	0.29 (0.026)	-3%	-2%

1. Relative to complete case analysis
2. FMI: Fraction of missing information

Table 5-12: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when IQ MNAR: difference in Pr(IQ observed)=0.10 for 1 SD increase in IQ; 40% missing data

		Complete case		MI				IPW			FIML		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]
IQ 40% missing	β_4	0.07 (0.041)	-29%	0.07 (0.041)	-28%	0%	45%	0.07 (0.041)	-29%	0%	0.07 (0.039)	-27%	10%
Correlation	β_5	0.16 (0.035)	-20%	0.16 (0.035)	-20%	-1%	41%	0.16 (0.035)	-20%	-2%	0.16 (0.035)	-19%	-2%
(IQ:KS4) = 0.1	β_6	0.26 (0.029)	-15%	0.26 (0.029)	-14%	0%	42%	0.26 (0.030)	-15%	-3%	0.26 (0.029)	-14%	3%
IQ 40% missing	β_4	As above		0.07 (0.040)	-26%	4%	43%	0.07 (0.041)	-27%	-1%	0.08 (0.039)	-25%	13%
Correlation	β_5			0.16 (0.035)	-18%	2%	40%	0.16 (0.035)	-18%	-2%	0.16 (0.035)	-18%	1%
(IQ:KS4) = 0.3	β_6			0.26 (0.029)	-14%	3%	40%	0.26 (0.030)	-14%	-3%	0.26 (0.028)	-13%	6%
IQ 40% missing	β_4	As above		0.08 (0.039)	-23%	12%	39%	0.08 (0.041	-23%	-1%	0.08 (0.037)	-21%	27%
Correlation	β_5			0.17 (0.034)	-16%	8%	36%	0.17 (0.035)	-15%	-2%	0.17 (0.034)	-15%	9%
(IQ:KS4) = 0.5	β_6			0.27 (0.028)	-12%	9%	37%	0.27 (0.030)	-11%	-3%	0.26 (0.028)	-12%	8%
IQ 40% missing	β_4	As above		0.08 (0.037)	-17%	27%	32%	0.08 (0.042)	-16%	-4%	0.09 (0.035)	-15%	39%
Correlation	β_5			0.18 (0.032)	-11%	20%	28%	0.18 (0.036)	-9%	-3%	0.18 (0.032)	-11%	20%
(IQ:KS4) = 0.7	β_6			0.28 (0.027)	-8%	22%	29%	0.28 (0.030)	-6%	-4%	0.27 (0.027)	-8%	18%
IQ 40% missing	β_4	As above		0.09 (0.032)	-7%	56%	16%	0.09 (0.043)	-5%	-8%	0.09 (0.032)	-6%	65%
Correlation	β_5			0.19 (0.029)	-5%	44%	14%	0.20 (0.037)	-1%	-7%	0.19 (0.029)	-4%	44%
(IQ:KS4) = 0.9	β_6			0.29 (0.025)	-3%	50%	14%	0.30 (0.031)	1%	-8%	0.29 (0.025)	-4%	40%

1. Relative to complete case analysis
2. FMI: Fraction of missing information

Table 5-13: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when IQ MNAR: difference in Pr(IQ observed)=0.10 for 1 SD increase in IQ; 60% missing data

		Complete case		MI				IPW			FIML		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]
IQ 60% missing	β_4	0.03 (0.049)	-74%	0.03 (0.050)	-73%	-1%	65%	0.02 (0.050)	-78%	-4%	0.02 (0.049)	-75%	0%
Correlation	β_5	0.10 (0.043)	-49%	0.10 (0.043)	-49%	-1%	62%	0.10 (0.043)	-51%	-1%	0.10 (0.043)	-48%	-2%
(IQ:KS4) = 0.1	β_6	0.19 (0.037)	-36%	0.19 (0.037)	-35%	-1%	63%	0.19 (0.037)	-37%	-1%	0.20 (0.036)	-35%	3%
IQ 60% missing	β_4	As above		0.03 (0.049)	-68%	1%	64%	0.02 (0.050)	-77%	-4%	0.03 (0.048)	-70%	4%
Correlation	β_5			0.11 (0.042)	-46%	1%	61%	0.10 (0.043)	-50%	-2%	0.11 (0.042)	-45%	3%
(IQ:KS4) = 0.3	β_6			0.20 (0.037)	-33%	2%	61%	0.19 (0.037)	-36%	-1%	0.20 (0.036)	-33%	7%
IQ 60% missing	β_4	As above		0.04 (0.047)	-58%	8%	60%	0.03 (0.050)	-75%	-4%	0.04 (0.047)	-60%	11%
Correlation	β_5			0.12 (0.041)	-39%	11%	57%	0.10 (0.044)	-48%	-3%	0.12 (0.041)	-39%	7%
(IQ:KS4) = 0.5	β_6			0.22 (0.035)	-28%	12%	57%	0.20 (0.038)	-34%	-2%	0.21 (0.034)	-28%	17%
IQ 60% missing	β_4	As above		0.06 (0.044)	-42%	28%	52%	0.03 (0.051)	-71%	-6%	0.06 (0.043)	-43%	34%
Correlation	β_5			0.14 (0.037)	-28%	31%	48%	0.11 (0.044)	-45%	-6%	0.14 (0.038)	-28%	29%
(IQ:KS4) = 0.7	β_6			0.24 (0.032)	-20%	32%	48%	0.21 (0.038)	-31%	-4%	0.24 (0.032)	-21%	36%
IQ 60% missing	β_4	As above		0.08 (0.036)	-17%	84%	30%	0.03 (0.052)	-67%	-10%	0.08 (0.035)	-17%	93%
Correlation	β_5			0.18 (0.032)	-11%	83%	27%	0.12 (0.045)	-40%	-12%	0.18 (0.032)	-12%	82%
(IQ:KS4) = 0.9	β_6			0.28 (0.027)	-8%	86%	28%	0.22 (0.039)	-27%	-10%	0.27 (0.027)	-8%	86%

1. Relative to complete case analysis
2. FMI: Fraction of missing information

Table 5-14: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when IQ MNAR: difference in Pr(IQ observed)=0.10 for 1 SD increase in IQ; 80% missing data

		Complete case		MI				IPW			FIML		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]
IQ 80% miss	β_4	-0.14 (0.068)	-237%	-0.14 (0.069)	-236%	-3%	86%	-0.18 (0.069)	-278%	-3%	-0.15 (0.069)	-247%	-3%
Correlation	β_5	-0.13 (0.062)	-165%	-0.13 (0.062)	-164%	-1%	85%	-0.18 (0.061)	-191%	1%	-0.13 (0.063)	-163%	-4%
(IQ:KS4) = 0.1	β_6	-0.05 (0.052)	-116%	-0.05 (0.053)	-115%	-1%	85%	-0.09 (0.052)	-130%	1%	-0.05 (0.053)	-117%	-3%
IQ 80% miss	β_4	As above		-0.12 (0.068)	-223%	0%	85%	-0.18 (0.069)	-279%	-3%	-0.13 (0.068)	-233%	1%
Correlation	β_5			-0.11 (0.060)	-155%	6%	84%	-0.18 (0.061)	-192%	1%	-0.11 (0.062)	-153%	1%
(IQ:KS4) = 0.3	β_6			-0.03 (0.051)	-109%	5%	84%	-0.09 (0.052)	-131%	1%	-0.03 (0.052)	-110%	0%
IQ 80% miss	β_4	As above		-0.09 (0.065)	-194%	9%	84%	-0.18 (0.069)	-282%	-3%	-0.10 (0.061)	-202%	25%
Correlation	β_5			-0.07 (0.056)	-134%	21%	81%	-0.19 (0.062)	-194%	0%	-0.07 (0.056)	-134%	20%
(IQ:KS4) = 0.5	β_6			0.02 (0.048)	-94%	19%	82%	-0.10 (0.052)	-132%	-1%	0.01 (0.048)	-96%	20%
IQ 80% miss	β_4	As above		-0.04 (0.059)	-143%	36%	79%	-0.19 (0.070)	-287%	-4%	-0.05 (0.056)	-149%	50%
Correlation	β_5			0.002 (0.050)	-99%	56%	76%	-0.20 (0.062)	-198%	-2%	0.004 (0.051)	-98%	48%
(IQ:KS4) = 0.7	β_6			0.09 (0.043)	-70%	50%	77%	-0.10 (0.053)	-134%	-4%	0.09 (0.043)	-71%	45%
IQ 80% miss	β_4	As above		0.04 (0.044)	-58%	140%	59%	-0.19 (0.071)	-295%	-7%	0.04 (0.044)	-62%	139%
Correlation	β_5			0.12 (0.038)	-41%	170%	55%	-0.21 (0.064)	-204%	-8%	0.12 (0.039)	-40%	148%
(IQ: KS4) = 0.9	β_6			0.21 (0.033)	-29%	156%	57%	-0.12 (0.055)	-138%	-11%	0.21 (0.034)	-30%	132%

1. Relative to complete case analysis
2. FMI: Fraction of missing information

5.2.4.4 Outcome MNAR, Factor 4 (interaction)

When an interaction was introduced between the exposure and outcome with respect to the probability of the outcome being observed (such that the probability that IQ was observed was more strongly related to IQ itself among those who had not been breastfed compared to those who had) the bias in all coefficients was exacerbated, particularly at higher levels of missing data. (The results for a correlation of 0.7 between IQ and the linked attainment variable are shown in Table 5-15; results for all correlations are given in Appendix B, Tables 8 to 11). Nonetheless, the bias was always reduced – and precision increased – through MI (and FIML) incorporating the linked variable as an auxiliary variable. The MI and FIML results were very similar.

The IPW results when the interaction (between breastfeeding and attainment) was included in the model to generate the weights were generally quite similar (on the whole, they were slightly more biased but the differences were small, particularly for 20% and 40% missing data) to the estimates obtained using MI (and FIML) except when there was 80% data. At 80% missing data, the IPW estimates were more biased than the MI/FIML and complete case estimates for all correlations between IQ and attainment. For all levels of missing data, IPW resulted in gains in efficiency (except when the correlation between IQ and attainment was 0.1), but these gains were slightly lower than those obtained using MI for correlations of 0.5, 0.7 and 0.9 (and similar when the correlation was 0.3).

When the IPW model excluded this interaction, the results were similar (to when the interaction was included in the model used to generate the weights) when there was a low correlation between IQ and attainment. For all other scenarios the results were more biased, particularly for high levels of missing data and when the correlation between IQ and attainment was high. These results are shown in Appendix B, Tables 8 to 11.

Table 5-15: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when IQ MNAR with an interaction[1] between breastfeeding and IQ with respect to the probability of IQ being observed (Factor 4)

		Complete case		MI				IPW[2]			FIML		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[3]	FMI[4]	Estimate (empirical SE)	% bias	Gain in precis-ion[3]	Estimate (empirical SE)	% bias	Gain in precis-ion[3]
IQ 20% missing	β_4	0.06 (0.034)	-42%	0.08 (0.032)	-24%	15%	15%	0.07 (0.032)	-27%	13%	0.08 (0.032)	-24%	16%
Correlation	β_5	0.13 (0.031)	-36%	0.16 (0.029)	-20%	12%	13%	0.16 (0.029)	-23%	10%	0.16 (0.029)	-19%	10%
(IQ:KS4) = 0.7	β_6	0.21 (0.026)	-31%	0.25 (0.025)	-17%	11%	13%	0.24 (0.025)	-20%	8%	0.25 (0.025)	-17%	8%
IQ 40% missing	β_4	0.03 (0.040)	-66%	0.06 (0.036)	-37%	23%	31%	0.06 (0.037)	-39%	18%	0.06 (0.037)	-35%	14%
Correlation	β_5	0.10 (0.036)	-51%	0.14 (0.032)	-29%	24%	28%	0.14 (0.033)	-25%	19%	0.14 (0.033)	-30%	18%
(IQ: KS4) = 0.7	β_6	0.17 (0.030)	-44%	0.23 (0.027)	-25%	22%	29%	0.22 (0.028)	-25%	16%	0.22 (0.027)	-25%	25%
IQ 60% missing	β_4	-0.03 (0.048)	-132%	0.02 (0.042)	-76%	30%	51%	0.01 (0.043)	-88%	23%	0.02 (0.042)	-77%	29%
Correlation	β_5	0.003 (0.043)	-99%	0.09 (0.038)	-56%	30%	48%	0.08 (0.039)	-61%	23%	0.09 (0.038)	-55%	29%
(IQ: KS4) = 0.7	β_6	0.06 (0.037)	-79%	0.17 (0.032)	-45%	38%	49%	0.16 (0.033)	-46%	28%	0.17 (0.031)	-45%	47%
IQ 80% missing	β_4	-0.23 (0.070)	-330%	-0.09 (0.058)	-190%	47%	77%	-0.20 (0.062)	-299%	29%	-0.09 (0.059)	-195%	40%
Correlation	β_5	-0.28 (0.062)	-242%	-0.08 (0.052)	-140%	41%	74%	-0.21 (0.055)	-203%	24%	-0.08 (0.049)	-140%	55%
(IQ: KS4) = 0.7	β_6	-0.25 (0.054)	-184%	-0.02 (0.045)	-106%	43%	76%	-0.12 (0.049)	-139%	18%	-0.02 (0.045)	-107%	42%

1. Difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when exposure=0 (no breastfeeding); change in difference in Pr(IQ observed) for each 1 SD increase in IQ = -0.025 for each increase in breastfeeding category (Factor 4 in scenarios)
2. Interaction term (between breastfeeding and linked attainment score) included in logistic model used to generate weights
3. Relative to complete case analysis
4. FMI: Fraction of missing information

5.2.4.5 Outcome MNAR, Factor 5 (linked attainment score MNAR)

Tables 5-16 and 5-17 show the results when missingness was introduced in the linked attainment score. When the association between the linked variable and the probability of it being observed was in the opposite direction to the relationship between IQ and the probability of IQ being observed, the estimates obtained from MI and FIML were very similar to those obtained with no missing data for the linked attainment score. When the association was in the same direction as that for IQ, the estimates from FIML were again very similar to those obtained with no missing data for the linked attainment score whereas the MI results were slightly more biased, except when there was 80% missing data. However, the differences were quite small and these estimates were still substantially less biased than those obtained from the complete case analysis. The IPW estimates with missing data in the linked attainment score were also quite similar to those obtained with no missing data; however, in contrast to the MI results, when missingness in the linked attainment score was in the same direction as missingness in IQ, the bias was slightly lower than when it was in the opposite direction. Missingness in the linked attainment score resulted in a loss of precision in the IPW estimates.

Table 5-16: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when linked attainment score was MNAR with 20% missing data (Factor 5); 20% and 40% missing data

[different values for difference in Pr(KS4 observed) for one SD increase in KS4; Pr(IQ observed)=0.10 for 1 SD increase in IQ)]

	Estimand	Complete case[1]		MI				IPW			FIML		
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[2]	FMI[3]	Estimate (empirical SE)	% bias	Gain in precis-ion[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[2]
Outcome 20% missing													
Diff	β_4	0.08 (0.034)	-17%	0.09 (0.033)	-7%	8%	17%	0.10 (0.039)	3%	-24%	0.09 (0.033)	-9%	6%
$Pr(KS4_{obs})$[4] =	β_5	0.17 (0.030)	-14%	0.19 (0.030)	-7%	6%	14%	0.20 (0.035)	1%	-23%	0.19 (0.029)	-8%	13%
-0.10	β_6	0.26 (0.025)	-12%	0.28 (0.026)	-6%	7%	15%	0.30 (0.028)	-1%	-18%	0.28 (0.024)	-6%	19%
Diff	β_4	As above		0.09 (0.033)	-12%	7%	18%	0.11 (0.038)	6%	-22%	0.09 (0.032)	-4%	13%
$Pr(KS4_{obs})$ =	β_5			0.18 (0.030)	-11%	7%	15%	0.20 (0.035)	-1%	-15%	0.18 (0.030)	-7%	8%
+0.10	β_6			0.27 (0.025)	-9%	9%	15%	0.30 (0.030)	-1%	-19%	0.28 (0.025)	-7%	9%
Outcome 40% missing													
Diff	β_4	0.07 (0.041)	-29%	0.09 (0.037)	-15%	16%	34%	0.08 (0.046)	-16%	-19%	0.09 (0.037)	-14%	16%
$Pr(KS4_{obs})$ =	β_5	0.16 (0.035)	-20%	0.18 (0.034)	-10%	16%	31%	0.18 (0.041)	-12%	-19%	0.18 (0.032)	-11%	31%
-0.10	β_6	0.26 (0.029)	-15%	0.28 (0.027)	-8%	17%	32%	0.28 (0.035)	-7%	-24%	0.28 (0.028)	-8%	17%
Diff	β_4	As above		0.08 (0.037)	-20%	22%	35%	0.09 (0.043)	-13%	-18%	0.08 (0.036)	-17%	30%
$Pr(KS4_{obs})$ =	β_5			0.17 (0.034)	-15%	15%	32%	0.18 (0.040)	-9%	-21%	0.18 (0.032)	-11%	26%
+0.10	β_6			0.27 (0.028)	-11%	13%	32%	0.28 (0.033)	-6%	-16%	0.28 (0.027)	-8%	23%

1. The results for the complete case analysis are the same as those presented in Tables 5-10 and 5-11 but are included here for comparison
2. Relative to complete case analysis
3. FMI: Fraction of missing information
4. Diff $Pr(KS4_{obs})$: difference in the probability that KS4 (attainment) was observed for 1 SD increase in KS4

Table 5-17: Estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) when linked attainment score was MNAR with 20% missing data (Factor 5); 60% and 80% missing data

[different values for difference in Pr(KS4 observed) for one SD increase in KS4; Pr(IQ observed)=0.10 for 1 SD increase in IQ]

	Estimand	Complete case[1]		MI				IPW			FIML		
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[2]	FMI[3]	Estimate (empirical SE)	% bias	Gain in precis-ion[2]	Estimate (empirical SE)	% bias	Gain in precis-ion[2]
Outcome 60% missing													
Diff	β_4	0.03 (0.049)	-74%	0.06 (0.043)	-43%	32%	55%	0.02 (0.056)	-76%	-26%	0.06 (0.041)	-42%	45%
$Pr(KS4_{obs})$[4] =	β_5	0.10 (0.043)	-49%	0.14 (0.039)	-30%	24%	52%	0.10 (0.048)	-50%	-23%	0.15 (0.037)	-27%	35%
-0.10	β_6	0.19 (0.037)	-36%	0.24 (0.032)	-21%	29%	52%	0.20 (0.043)	-33%	-26%	0.24 (0.031)	-20%	38%
Diff	β_4	As above		0.05 (0.042)	-50%	24%	55%	0.03 (0.056)	-66%	-26%	0.06 (0.042)	-43%	28%
$Pr(KS4_{obs})$ =	β_5			0.13 (0.037)	-35%	35%	52%	0.11 (0.051)	-43%	-24%	0.14 (0.037)	-29%	32%
+0.10	β_6			0.23 (0.031)	-25%	33%	52%	0.21 (0.041)	-30%	-22%	0.24 (0.032)	-23%	21%
Outcome 80% missing													
Diff	β_4	-0.14 (0.068)	-237%	-0.06 (0.060)	-162%	28%	81%	-0.20 (0.079)	-300%	-24%	-0.05 (0.058)	-146%	38%
$Pr(KS4_{obs})$ =	β_5	-0.13 (0.062)	-165%	-0.02 (0.054)	-111%	30%	78%	-0.21 (0.070)	-207%	-26%	0.005 (0.052)	-97%	40%
-0.10	β_6	-0.05 (0.052)	-116%	0.07 (0.047)	-78%	25%	80%	-0.13 (0.061)	-142%	-28%	0.09 (0.045)	-70%	39%
Diff	β_4	As above		-0.06 (0.060)	-156%	37%	80%	-0.18 (0.078)	-283%	-22%	-0.05 (0.057)	-148%	52%
$Pr(KS4_{obs})$ =	β_5			-0.01 (0.053)	-108%	36%	77%	-0.18 (0.068)	-190%	-15%	0.004 (0.051)	-98%	50%
+0.10	β_6			0.07 (0.045)	-76%	30%	79%	-0.09 (0.057)	-129%	-19%	0.09 (0.042)	-70%	49%

1. The results for the complete case analysis are the same as those presented in Tables 5-12 and 5-13 but are included here for comparison
2. Relative to complete case analysis
3. FMI: Fraction of missing information
4. Diff $Pr(KS4_{obs})$: difference in the probability that KS4 (attainment) was observed for 1 SD increase in KS4

5.2.4.6 Sensitivity analysis

As mentioned on page 105, in a sensitivity analysis I made attainment dependent on sex and mother's education in addition to IQ. A limited set of scenarios were investigated using multiple imputation only (the complete case analysis would be unaffected). These results are given in Appendix B, Table 12. The results were essentially the same as when attainment was only dependent on IQ.

5.3 Discussion

Through the simulations I show that, if a continuous outcome variable is MNAR, the estimate of the exposure-outcome association obtained from a complete case analysis will be biased (if there is a causal association between the exposure and outcome). Further, bias will increase as the percent of missing data increases. In the absence of linked data, IPW, MI and FIML are likely to produce estimates that are similarly biased. On the other hand, having a linked proxy (or several proxies) for a missing continuous outcome variable and including it as an auxiliary variable in MI or FIML will result in reductions in bias as long as the correlation between the proxy and the missing study variable is reasonably high (>0.3); there may also be substantial gains in efficiency. Using IPW (with the linked variable or variables as additional predictors of the probability of missingness) could result in greater reductions in bias (i.e. greater than those seen using MI or FIML) but is more likely to lead to smaller reductions in bias, particularly at higher levels of missing data. IPW will also result in a loss of efficiency, particularly if there is missingness in the linked variable(s).

In term of the exemplar presented in this chapter, the univariate correlation between the capped KS4 attainment score and IQ was 0.6 and, jointly, the transformed KS2, KS3 and KS4 attainment variables explained 44% of the variability in IQ (coefficient of multiple correlation = 0.66). Additionally, 64% of individuals had missing data on IQ (although the percentage with any missing data was higher – only 26% of individuals were complete cases). As such, I would expect the MI (and FIML) estimates for this exemplar to be less biased than the estimates obtained from the complete case

analysis, although they are likely to still be under-estimates of the true effect of breastfeeding on IQ (assuming no residual confounding). Thus, in the absence of residual confounding, the estimated effect of increased duration of breastfeeding on mean IQ is likely to be higher than the (adjusted) estimates given in the last two columns of Table 5-7.

The IPW estimates in the exemplar were quite similar to those obtained from MI, particularly after truncating large weights whereas, in the simulations, the IPW estimates with large amounts of missing data were quite similar to the complete case estimates. The simulations I carried out did not exactly match the ALSPAC data. I only simulated missing outcome data whereas, in fact, there was also missing data in the exposure and other covariates; in addition, the simulated datasets only contained breastfeeding and two covariates. Further, I simulated missingness in IQ such that the probability of missingness decreased linearly as IQ increased. Although the probability of IQ being missing did appear to decrease in a reasonably linear fashion across the deciles of attainment (these results not shown), it is not possible to determine whether the same would apply with IQ. These differences could explain why the IPW estimates in the exemplar were not more different from those obtained using MI.

In summary, if a continuous outcome variable is MNAR, including proxies for this outcome obtained from linked datasets as auxiliary variables in multiple imputation (or FIML) models will reduce bias and increase efficiency under a wide range of conditions, even with high levels of missing data, particularly when these proxies have reasonably high correlations – either individually or jointly – with the study outcome. Use of IPW may reduce or increase bias and is likely to lead to a loss of precision.

Chapter 6 Missing binary outcome

In Chapter 5 I examined missingness in a continuous outcome; this chapter investigates the impact of missingness in a binary outcome, again focussing on the situation when the outcome is likely to be MNAR. As previously, I do this through an exemplar and a simulation study. As outlined in Chapters 1 and 2, with a binary outcome a complete case analysis using logistic regression will produce an unbiased estimate of the odds ratio for exposure if there is no multiplicative interaction between the exposure and outcome with respect to the probability of missingness (Bartlett et al. 2015). In contrast, IPW would be expected to result in biased estimates of the exposure odds ratio if missingness depended on the binary outcome. If there were only missingness in the binary outcome and, as would be the default, this were imputed using a logistic model, you would expect MI to give (asymptotically) unbiased estimates. If, however, there were covariates that were MNAR, MI may result in bias (since these would be being imputed assuming MAR).

In this Chapter I will therefore be using linked data to examine the likely missingness mechanism in order to inform decisions about the most appropriate analysis strategy in the exemplar. I will then compare the different analysis approaches, both in the exemplar and through the simulations.

Some of this work presented in this chapter has been previously published: Cornish RP et al. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. BMJ Open 2016 6:e013167. https://doi.org/10.1136/bmjopen-2016-013167

6.1 Exemplar: maternal smoking in pregnancy and offspring depression

The background to this exemplar is given in Chapter 2 (Section 2.5.2). Briefly, I am interested in whether maternal smoking during pregnancy is associated with an increased odds of offspring depression (at age 18 years). I use linked data on GP-recorded depression to investigate missingness in ALSPAC-measured depression and as auxiliary variables in the MI and IPW analyses. The analysis model for this exemplar is a logistic regression of depression on smoking in pregnancy and the covariates listed in Table 6-1.

6.1.1 Analysis

Subjects included in this analysis were all singletons and twins enrolled in the study who were alive at one year, had not withdrawn or explicitly dissented to linkage to their health records and for whom ALSPAC had a valid NHS ID number (n=14,566).

6.1.1.1 Variables

The variables used are described in detail in Chapter 3, so are simply listed in Table 6-1.

Table 6-1: Variables included in the analysis of Exemplar 2

Variable	Details	Chapter 3 section
Outcome	(Offspring) depression at 18 years (meets ICD-10 criteria for a diagnosis - yes/no)	3.3.1
Exposure	Smoking during pregnancy (yes/no) – reported at 18 and 32 weeks gestation and when the baby was 8 weeks old.	3.3.2
Covariates	Child's sex; maternal age (at the child's birth and at first pregnancy), parity, marital status, ethnicity, alcohol and drug use in early pregnancy; maternal and paternal depression and anxiety during pregnancy; maternal and paternal education; family occupational social class; housing tenure; use of car; phone in home; double glazing; number of rooms in the house and financial difficulties score.	3.3.3
Auxiliary variables from ALSPAC	Maternal smoking (ever smoked), reported at 18 weeks gestation; emotional difficulties score from the Strength and Difficulties Questionnaire (SDQ) at age 7.	3.3.3
Linked GP variables	Current diagnosis or symptoms or treatment, historical diagnosis or symptoms or treatment, future diagnosis or symptoms or treatment (see below in Section 6.1.1.1.1).	3.4.2.1

As stated in Chapter 5, mother's age at first pregnancy was strongly associated with maternal age which resulted in the MI models not converging. Although there was no evidence that age at first pregnancy was associated with either depression or missingness in depression in the current analysis, it was strongly associated with missingness in the GP measures of depression. Therefore, this variable was omitted from the complete case analysis but included in the IPW and MI models that included the linked GP data. However, so that the MI models would converge, some of the maternal age and age at first pregnancy categories were combined. I used the new maternal age variable in the complete case analysis so that the results from this and the MI models were comparable (i.e. adjusted for identical variables). Similarly, since I also found that father's education was associated with missingness in the GP data (Chapter 4), this was also included in the MI and IPW models that included the linked GP data but not in the complete case analysis.

6.1.1.1.1 Linked data

In Chapter 3 (Section 3.4.2.1) I described the code lists used previously to define depression using GP data. In addition to these, I included a depression symptom code for "Loss of interest". For this analysis I created nine different outcomes: current diagnosis, symptoms and treatment for depression (current = 6 months either side of the month in which the CIS-R was completed – if completed – and between the ages of 17 years and 4 months and 18 years and 4 months if not completed, since the average age at completion was 17 years and 10 months); historical diagnosis symptoms and treatment (historical = more than 6 months before the month in which the CIS-R was completed / before age 17 years 4 months if not completed); and future diagnosis, symptoms and treatment (future = more than 6 months after the month of CIS-R completion / after age 18 years 4 months if not completed). In addition, I used these separate variables to create three combined measures: current diagnosis or symptoms or treatment, historical diagnosis or symptoms or treatment, and future diagnosis or symptoms or treatment.

6.1.1.2 Relationship between ALSPAC-recorded and GP-recorded depression

A series of 2 x 2 tables were used to examine the level of agreement between CIS-R defined depression (as measured in ALSPAC) and GP-recorded depression. Logistic regression was used to assess which of the GP-recorded measures independently predicted CIS-R defined depression.

6.1.1.3 Examining the missing data mechanism

As in the previous chapter, logistic regression was used to examine the predictors of missingness. The factors included were those identified in Chapter 4 (excluding age at first pregnancy, as explained above); the linked measures of depression were included to investigate whether there was evidence for CIS-R defined depression being MNAR.

6.1.1.4 Dealing with missing data

For this exemplar, three different approaches were used to deal with missing data when modelling the relationship between smoking in pregnancy and offspring depression:

a) A complete case analysis
b) Inverse probability weighting (IPW), both including and excluding the linked GP measures of depression.
c) Multiple imputation using chained equations, also performed both with and without the linked variables.

6.1.1.4.1 IPW

This method is described in Section 2.1.3. I used four (logistic) models to obtain the inverse probability weights. All four models included the baseline covariates (including the exposure, smoking in pregnancy). In addition, each model included one or more auxiliary variables. There were two models that did not include the linked data. In model 1, only maternal smoking (ever smoked, asked at 18 weeks gestation) was included; in model 2, both maternal smoking (ever smoked) and the emotional symptoms score from the SDQ at age 7 years were added. Finally, the three composite measures of depression (historical, current and future diagnosis or treatment or symptoms) were added to each of these two models, giving models 3 and 4. This was done to gauge the separate effects of adding these different types of auxiliary variable. The IPW models using auxiliary variables included only those with fully observed covariates and relevant auxiliary variable(s). The Hosmer-Lemeshow goodness of fit test (Hosmer 1989) was used to assess the fit of the logistic models used to generate the (inverse probability) weights. As a sensitivity analysis, large weights were truncated – choosing different maximum values (8, 6 and 4).

6.1.1.4.2 MI models

I used four MI models; these included all the variables in the analysis model. The four MI models included the same auxiliary variables as the four IPW models: model 1 included maternal smoking (ever smoked), model 2 included maternal smoking and

emotional difficulties score at age 7 and, as above, the three GP measures of depression were added to each of these two models, giving models 3 and 4. In all cases, 100 datasets were imputed using 10 burn-in iterations. Since the SDQ emotional symptoms score as well as the maternal and paternal anxiety and depression scores were skewed, these were imputed using predictive mean matching (Morris et al. 2014), sampling from a group of ten potential matches. All other variables were imputed using the default regression model for that type of variable (ordinary least squares regression for continuous variables, logistic regression for binary variables, and multinomial logistic regression for categorical variables.

6.1.2 Results

There were 14,684 singletons and twins alive at one year who had not subsequently withdrawn from the study. Of these, ALSPAC had no NHS number for 118 individuals. Thus, this analysis is based on the remaining 14,566 individuals. Of these, 11,227 (77%) had data on maternal smoking in pregnancy, 4,537 (31%) had offspring depression data, 3,733 (26%) had non-missing data on both the exposure and outcome and 2,471 (17%) were complete cases (individuals with smoking in pregnancy, depression, and complete covariate information, but not necessarily linked GP data). In addition, among the 14,566 individuals included in this analysis, linked GP data were available for 10,560 (72%). Of these, 9,090 had the GP data necessary to determine whether they had current depression (6 months either side of the date at which they attended the ALSPAC clinic when the CIS-R was completed); this equated to 62% of the original 14,566. Further details of the available data are given in Table 6-2.

Table 6-2: Completeness of ALSPAC data by availability of GP data

Complete data on:			Linked GP data		Total
Covariates	Smoking in pregnancy	Depression (CIS-R)	Yes[1]	No	
Yes	Yes	Yes	2,011 (81%)	460 (19%)	2,471
		No	2,546 (68%)	1,214 (32%)	3,760
	No	Yes	137 (85%)	25 (15%)	162
		No	202 (68%)	99 (32%)	299
No	Yes	Yes	1,020 (81%)	242 (19%)	1,262
		No	2,573 (69%)	1,161 (31%)	3,734
	No	Yes	521 (81%)	121 (19%)	642
		No	1,550 (69%)	686 (31%)	2,236
			10,560 (72%)	4,006 (28%)	14,566

1. Information on at least one of: historical, current or future diagnosis or treatment or symptoms of depression (whether positive or negative)

6.1.2.1 Association between CIS-R and GP-recorded depression

Table 6-3 shows the relationship between the various depression outcomes defined using the GP data and CIS-R defined depression. Most individuals (over 98.5%) who did not meet the ICD-10 criteria for a diagnosis of depression using the CIS-R did not have a current record of diagnosis, treatment, or symptoms of depression in their GP data. However, only just over a quarter of individuals with depression as measured using the CIS-R had a current diagnosis, treatment or symptoms (Table 6-3). The results were similar for historical diagnoses, symptoms and treatment. Future diagnosis, symptoms and treatment all had higher sensitivities but lower specificities than either current or historical. Just over 55% of individuals with CIS-R defined depression had a future diagnosis, symptoms or treatment.

After mutual adjustment, the combined factors were all strongly associated with CIS-R defined depression: current diagnosis or symptoms or treatment OR = 5.04 (95% CI: 3.11, 8.17); future diagnosis or symptoms or treatment OR=3.14 (2.37, 4.17); and historical diagnosis or symptoms or treatment OR=2.31 (1.44, 3.69).

Table 6-3: ALSPAC-measured (CIS-R) depression by GP measures of depression

		CIS-R diagnosis of depression	
Case definition[1]		No	Yes
Current treatment	No	3038 (98.5%)	226
	Yes	45	40 (15.0%)
Current diagnosis	No	3059 (99.4%)	247
	Yes	18	16 (6.1%)
Current symptoms	No	3040 (98.7%)	220
	Yes	39	43 (16.1%)
Current diagnosis or symptoms or treatment	No	3012 (97.7%)	199
	Yes	72	71 (26.3%)
Future treatment	No	2605 (83.3%)	135
	Yes	524	146 (52.0%)
Future diagnosis	No	2900 (93.7%)	208
	Yes	195	62 (23.0%)
Future symptoms	No	2789 (89.7%)	186
	Yes	321	87 (31.9%)
Future diagnosis or symptoms or treatment	No	2500 (79.6%)	126
	Yes	640	156 (55.3%)
Historical treatment	No	3300 (98.2%)	245
	Yes	60	36 (12.8%)
Historical diagnosis	No	3335 (99.3%)	265
	Yes	25	16 (5.7%)
Historical symptoms	No	3288 (97.9%)	243
	Yes	72	38 (13.5%)
Historical diagnosis or symptoms or treatment	No	3233 (96.2%)	217
	Yes	127	64 (22.8%)

1. The denominators vary because the numbers with historical, current and future data on depression are different

6.1.2.2 Predictors of missing ALSPAC-measured depression

Table 6-4 shows the prevalence of current, historical and future diagnosis, symptoms and treatment according to whether individuals completed the CIS-R in ALSPAC. The proportions with a current, future or historical diagnosis were very similar amongst those with and without CIS-R data. Those with missing CIS-R data were slightly more likely to have a current or historical record of symptoms or treatment. Likewise, those with missing CIS-R data were more likely to have a record of future symptoms and treatment; these differences were more marked than for current or historical diagnosis and treatment.

Table 6-4: Current, future and historical diagnosis, symptoms and treatment of depression by availability of ALSPAC-measured depression

	ALSPAC-measured (CIS-R) depression data available[1]	
	Yes (n<=3,641)	No (n<=6,716)
Current treatment	85 (2.5%)	222 (3.9%)
Current diagnosis	34 (1.0%)	60 (1.1%)
Current symptoms	82 (2.5%)	179 (3.1%)
Current diagnosis, treatment or symptoms	143 (4.3%)	333 (5.8%)
Future treatment	670 (19.7%)	1,616 (27.0%)
Future diagnosis	257 (7.6%)	437 (7.6%)
Future symptoms	408 (12.1%)	1,057 (18.0%)
Future diagnosis, treatment or symptoms	796 (23.3%)	1,871 (31.0%)
Historical treatment	96 (2.6%)	213 (3.3%)
Historical diagnosis	41 (1.1%)	72 (1.1%)
Historical symptoms	110 (3.0%)	263 (3.9%)
Historical diagnosis, treatment or symptoms	191 (5.3%)	432 (6.4%)

1. Analysis restricted to individuals with GP data up to 6 months past TF4 clinic and/or 220 months, as appropriate (current or future events), or those with data beyond at least age 10 (historical events)

Table 6-5, 6-6 and 6-7 show the predictors of missing CIS-R depression data among those with complete data on smoking in pregnancy, covariates, and linked GP data (n=3,971). After adjusting for covariates, only future diagnosis, symptoms or treatment was associated with missing depression data; individuals with a future diagnosis, symptoms or treatment were more likely to having missing CIS-R depression data (Table 6-7).

Table 6-5: Predictors of missingness in ALSPAC-measured depression: child and maternal covariates

(n=3,971 with complete covariate information plus complete linked data)

Factor	Level	OR (95% CI)[1]	p-value
Sex	Female vs male	0.60 (0.53, 0.69)	p<0.001
Ethnicity	Non-white vs white	1.22 (0.69, 2.17)	p=0.5
Smoking in pregnancy	Yes vs no	1.58 (1.32, 1.91)	p<0.001
Mother's education	O level/lower	1.00	
	A level	0.72 (0.61, 0.85)	
	Degree	0.61 (0.49, 0.76)	p<0.001
Duration of breastfeeding	Never/<1 month	1.00	
	1-2 months	0.66 (0.54, 0.82)	
	3-5 months	0.71 (0.58, 0.86)	
	6+ months	0.52 (0.44, 0.62)	p<0.001
Mother's age at birth	<20	2.96 (1.63, 5.38)	
	20-29	1.27 (1.09, 1.48)	
	30+	1.00	p<0.001
Parity	0	1.00	
	1	1.30 (1.12, 1.52)	
	2+	1.77 (1.44, 2.18)	p<0.001
Mother's marital status	Not married vs married	0.91 (0.74, 1.12)	p=0.4
Maternal depression	Per 1 point increase	0.99 (0.97, 1.01)	p=0.5
Maternal anxiety	Per 1 point increase	1.02 (0.99, 1.05)	p=0.3

1. Mutually adjusted for all covariates plus linked GP depression variables

Table 6-6: Predictors of missingness in ALSPAC-measured depression: paternal and family covariates

(n=3,971 with complete covariate information plus complete linked data)

Factor	Level	OR (95% CI)[1]	p-value
Paternal depression	Per 1 point increase	0.99 (0.97, 1.02)	p=0.6
Paternal anxiety	Per 1 point increase	0.99 (0.96, 1.02)	p=0.5
Housing tenure	Mortgaged /owned	1.00	
	Private rented	1.81 (1.26, 2.61)	
	Council/HA/other	1.18 (0.90, 1.55)	p=0.005
Number of rooms	Per 1 room increase	0.93 (0.87, 0.99)	p=0.02
Phone in home	Yes vs no	1.02 (0.72, 1.45)	p=0.9
Use of car	Yes vs no	0.86 (0.60, 1.24)	p=0.4
Double glazing	Full/partial vs none	0.93 (0.81, 1.06)	p=0.3
Financial difficulties	Per 1 unit increase	0.99 (0.97, 1.02)	p=0.6
Family occupational social class	Manual vs non-manual	1.37 (1.11, 1.69)	p=0.003

1. Mutually adjusted for all covariates plus linked GP depression variables

Table 6-7: Predictors of missingness in ALSPAC-measured depression: GP recorded depression

(n=3,971 with complete covariates plus complete linked data)

Variable		OR (95% CI)[1]	p-value
Current diagnosis or symptoms or treatment	Yes	1.04 (0.69, 1.55)	p=0.9
Historical diagnosis or symptoms or treatment	Yes	1.13 (0.80, 1.59)	p=0.5
Future diagnosis or symptoms or treatment	Yes	1.21 (1.02, 1.43)	p=0.03

1. Adjusted for all covariates and mutually adjusted

There was no evidence for a multiplicative interaction between smoking in pregnancy and current GP-recorded depression with respect to the probability of missingness in CIS-R depression data [risk ratio (RR) for interaction between smoking in pregnancy

and current diagnosis or symptoms or treatment = 1.12 (0.67, 1.87), p=0.7; and RR for interaction with future diagnosis or symptoms or treatment = 0.90 (0.69, 1.17), p=0.4, when added to a binomial regression model including a restricted set of covariates (sex, mother's education, mother's age, parity, housing tenure, occupational social class, and duration of breastfeeding)]. These covariates were selected on the basis of their strength of association with missing depression data; only a restricted set of covariates could be included because models including a large number of covariates did not converge.

6.1.2.3 Predictors of missingness in GP-recorded depression

As described in Chapter 4 (Section 4.3), males, individuals with more highly educated fathers, those whose mother was older at first pregnancy, those breastfed for longer, those whose mother smoked during pregnancy, those living in non-owned/mortgaged properties (as measured during pregnancy), and those whose family occupational social class was classified as non-manual were more likely to have missing GP data. After adjusting for these factors, there was no evidence that ALSPAC-measured depression was associated with missingness in GP-recorded depression [OR=0.91, 95% CI (0.64, 1.30)].

6.1.2.4 Relationship between smoking in pregnancy and offspring depression

Table 6-8 gives the odds ratios for depression comparing offspring of mothers who smoked during pregnancy to offspring of non-smokers obtained using the various analysis approaches. The adjusted odds ratio from the complete case analysis was 1.47 (95% CI: 0.95, 2.27). Inclusion of the auxiliary variables from ALSPAC reduced the adjusted odds ratios for both MI and IPW (compared to the complete case analysis). IPW resulted in a loss in precision whereas MI increase precision substantially.

When the linked variables were included the odds ratios from MI were reduced slightly further (compared to MI excluding the linked variables but including auxiliaries from ALSPAC) whereas the estimates from the IPW models were increased and became close to the complete case estimate. Again, the IPW estimates were less

precisely estimated than the complete case estimate, whereas MI resulted in gains in precision.

Truncating the weights to 4 in the IPW models resulted in a small reduction in the estimate of the odds ratio (OR reduced from 1.43 to 1.41 for the final IPW model presented in Table 6-8).

Table 6-8: Relationship between smoking in pregnancy and offspring depression: odds ratio estimates obtained from different analysis approaches

	Analysis approach								
	Complete case analysis (n=2,471)	Excluding linked data				Including linked data			
		IPW[3] (n=2,363)	MI[3] (n=14,566)	IPW[4] (n=2,112)	MI[4] (n=14,566)	IPW[3] (n=1,769)	MI[3] (n=14,566)	IPW[4] (n=1,587)	MI[4] (n=14,566)
Crude OR (95% CI)	1.91 (1.32, 2.76)	2.01 (1.27, 3.17)	1.82 (1.38, 2.39)	1.88 (1.21, 2.92)	1.86 (1.41, 2.45)	2.17 (1.27, 3.71)	1.79 (1.38, 2.31)	2.15 (1.27, 3.63)	1.85 (1.41, 2.42)
Adjusted[1] OR (95% CI)	1.47 (0.95, 2.27)	1.20 (0.75, 1.93)	1.35 (0.99, 1.84)	1.27 (0.78, 2.07)	1.37 (0.99, 1.90)	1.48 (0.87, 2.52)	1.29 (0.94, 1.77)	1.43 (0.83, 2.44)	1.34 (0.96, 1.89)
Gain in precision[2]	N/A	-5%	+100%	-10%	+82%	-24%	+94%	-26%	+68%

1. Adjusted for sex, ethnicity, mothers age, parity, housing tenure, marital status, family social class, financial difficulties, maternal education, maternal and paternal depression and anxiety, drug use in pregnancy, alcohol use in pregnancy, duration of breastfeeding, use of car & phone, double glazing, number of rooms in house
2. Variance (log OR) from complete case analysis / variance (log OR) from MI or IPW (as applicable), expressed as a percentage decrease/increase
3. Including only ever smoked as an auxiliary variable from ALSPAC
4. Including both auxiliary variables from ALSPAC (ever smoked and SDQ emotional difficulties score at age 7)

6.1.3 Discussion

As discussed in Chapters 1 and 2 and in the introduction at the beginning of this chapter, with a binary outcome a complete case logistic regression will produce an unbiased estimate of the exposure odds ratio as long as there is not a multiplicative interaction between the exposure and the outcome with respect to the probability of missingness. In contrast, if the data are MNAR then both MI and IPW will result in bias. In the exemplar described above, I used linked GP data on depression (as a proxy for ALSPAC-measured depression) to demonstrate that there was no evidence for such an interaction. As such, the complete case analysis should give an unbiased estimate of the association between smoking in pregnancy and offspring depression. However, this assumes that missingness is truly related to the binary measure of the outcome. In fact, it is likely that missingness – if related to the outcome at all – would vary according to levels of an underlying continuous measure of the outcome (for example, symptom severity in the case of depression) rather than being only associated with whether or not an individual meets the diagnostic threshold. It is not clear to what extent this might bias estimates of the odds ratio obtained via a complete case logistic regression.

There was some evidence that the outcome in this exemplar (ALSPAC-measured depression) might have been MNAR, since missingness was associated with one of the measures of GP-recorded depression. Inclusion of the linked measures of depression would give a closer approximation to MAR and, as such, MI and IPW models including the linked variables might be expected to be less biased than those not including these auxiliary variables.

In the exemplar the MI estimate of the odds ratio (including linked variables and other auxiliary variables from ALSPAC) was slightly lower than the estimate obtained from the complete case analysis. The IPW estimate when including all auxiliary variables was higher than the MI estimate but still slightly lower than the complete case estimate. It is not possible to determine from the data which estimate is likely to be the least biased.

6.2 Simulations

This was based on the above exemplar, the aim being to try to determine which analysis is likely to be most appropriate in this instance. In this simulation study I made missingness in the outcome dependent on an underlying continuous measure. I varied the extent of missing data, the degree to which the outcome was MNAR and the sensitivity and specificity of the linked binary measure in terms of predicting the missing (binary) study outcome.

6.2.1 Simulated datasets

The variables included in the simulations were the exposure variable (maternal smoking during pregnancy) and the outcome variable (depression) – both as a continuous measure and a derived binary measure (the outcome in the analysis model). I simulated a single proxy variable, GP-measured depression. As in the simulation study described in Chapter 5, I simulated 1,000 datasets of size 10,000.

The exposure (smoking in pregnancy) was drawn from a Bernoulli distribution with probability 0.25 (to roughly match the prevalence of smoking during pregnancy seen in ALSPAC). The (offspring) depression score was simulated as a standard normal variable, dependent on the exposure variable as shown in Equation 1.

$$\text{Depression score}_i = \mu + \omega \times m\_smoke_i + \varepsilon_i \qquad [1]$$

where m_smoke represents a dummy variable for smoking in pregnancy. The value of the coefficient ω was chosen to approximate the estimate obtained from the ALSPAC data ($\omega = 0.2$) and μ and ε were chosen to give the depression score a mean and variance of 0 and 1, respectively. A binary depression variable was created from this score such that the prevalence of depression was 7.5%. This was done by dichotomising the depression score at 1.44 (cut-off for the standard normal distribution to give 7.5% in the tail).

The analysis model is given by Equation 2:

$$\text{Logit}(\text{p\_depressed}_i) = \beta_0 + \beta_1 \times m\_smoke_i \quad [2]$$

where β_0 is the log odds of depression among offspring whose mothers did not smoke during pregnancy and β_1 is the log odds ratio for depression comparing offspring whose mothers smoked to those whose mothers did not. The coefficients used in Equation 1 were such that the true odds ratio for depression (e^{β_1}) was 1.5 (comparing children whose mother smoked to children whose mother did not smoke in pregnancy).

The (binary) linked outcome (GP-measured depression) was created to give different sensitivities and specificities in relation to the study's binary measure of depression. For individuals whose continuous depression score was below a z-score of 1.04 (corresponding to 85% of the distribution), the GP measure was set to zero; then, for individuals with no depression according to the study binary measure, but with depression scores above this threshold, the GP measure of depression was generated as a Bernoulli random variable with probability calculated to give particular specificities (probability 0.64 to give specificity 95% and 0.128 to give specificity 99%).

Finally, for those with depression according to the study binary measure, the GP measure of depression was again generated as a Bernoulli random variable – where the probability of being defined as depressed according to the GP measure increased as the study continuous depression score increased (Equation 3).

$$\text{Logit}(\text{p\_sens}_i) = \rho + \ln(4) \times \text{depression score}_i \quad [3]$$

Values of ρ were chosen to give particular sensitivities (0.25 and 0.90).

6.2.2 Simulating the missing data

As in the simulation study described in Chapter 5, I only induced missingness in the outcome variable. In scenarios where I assumed the continuous depression score was a measured variable (rather than an underlying latent constraint), this was missing

whenever the dichotomous depression measure was missing. The outcome was simulated as MNAR. This was done in two ways (given as Factor 5 below).

Missingness dependent on exposure and continuous outcome but not their interaction

Two different probabilities were generated using logistic regression, as shown in Equations 4 and 5:

$$\text{logit}(p_{1i}) = \alpha_1 + \delta \times \text{depression score}_i \quad [4]$$

$$\text{logit}(p_{2i}) = \alpha_2 + \gamma \times m\_smoke_i \quad [5]$$

From these, I created two Bernoulli random variables R_1 and R_2 (i.e. with probabilities p_1 and p_2). The outcome was classified as being observed if both R_1 and R_2 were equal to 1, and missing otherwise. The values of α_1 and α_2 were chosen using trial and improvement to give specific percentages of missing data (20%, 40%, 60%, 80%) and the values of δ chosen to vary the degree to which the outcome was MNAR. In all these scenarios γ was fixed to be ln(0.75).

Missingness dependent on exposure, continuous outcome and their interaction

In simulations in which there was an interaction between the exposure and outcome with respect to missingness, the probabilities were generated from the logistic model shown in Equation 6. As above, the values of α were chosen using trial and improvement in order to produce particular percentages of missing data (20, 40, 60, 80%). In these scenarios with an interaction, δ, γ and ω were fixed at ln(0.9), ln(0.7) and ln(1.1), respectively.

$$\begin{aligned}\text{logit}(\text{P(observed)}_i) \quad [6]\\ = \alpha + \delta \times \text{depression score}_i + \gamma \times m\_smoke_i\\ + \omega \times \text{depression score}_i \times m\_smoke_i\end{aligned}$$

6.2.3 Scenarios

The following five factors were varied in the simulations:

Factor 1: The percentage of missing outcome (depression) data: 20%, 40%, 60%, 80%.

Factor 2: The extent to which the outcome was MNAR: risk ratio for observing depression for a one SD increase in depression score = 0.90, 0.75.

Factor 3: The sensitivity of the GP measure in terms of identifying those with depression (according to the binary study outcome): 25%, 90%.

Factor 4: The specificity of the GP measure in terms of identifying those with depression: 95%, 99%.

Factor 5: Whether or not there was an interaction between the exposure and the continuous outcome with respect to missingness (described above in Section 6.2.2).

In the main set of scenarios (without the interaction) I simulated every possible combination of Factors 1-4, giving 32 scenarios. In a secondary set I included the interaction (Factor 5); in this I varied Factors 1, 3, and 4, giving an extra 16 scenarios. This is summarised in Table 6-9. For each scenario I simulated 1,000 datasets.

Table 6-9: Scenarios investigated in the simulations based on Exemplar 2

Scenarios investigated with no interaction between exposure and outcome with respect to missingness (each investigated at 20%, 40%, 60%, 80% missing data – Factor 1)				
Factor 2: OR_{obs} for 1 SD increase in outcome (depression score)	**Factor 3**: Sensitivity of GP measure for identifying those with depression			
	25%		90%	
	Factor 4: Specificity of GP measure			
	95%	99%	95%	99%
0.75	✓	✓	✓	✓
0.90	✓	✓	✓	✓
Scenarios investigated with interaction (each investigated at 20%, 40%, 60%, 80% missing data – Factor 1)				
0.90[1]	✓	✓	✓	✓

1. With OR_{obs} (odds ratio for being observed) for interaction between exposure and outcome = 1.1

6.2.4 Statistical analysis

I estimated the log odds ratio for smoking in pregnancy using logistic regression with smoking in pregnancy as the only covariate. This was estimated using:

a) A complete case analysis
b) Inverse probability weighting. The logistic regression model used to generate the inverse probability weights included the exposure and the linked depression variable. In addition, when the interaction (between the exposure and outcome) was introduced in the missingness model, the model used to generate the weights included an interaction between the exposure and the linked depression variable. Note that, as in Chapter 5, I also carried out the IPW with this interaction omitted from the model.
c) Multiple imputation. For each simulated dataset, 100 imputed datasets were created. Two different imputation models were used; both included the exposure (smoking in pregnancy) and the linked depression variable (binary). In the first model, the continuous outcome was imputed and the binary outcome imputed passively (by dichotomising the imputed continuous outcome at the cut-off described above, 1.44). In the second model, the continuous outcome was omitted and the binary outcome imputed directly.

As in Chapter 5, the estimates obtained from these analyses were compared to the true parameter (log odds ratio = 0.402). For this log odds ratio (β_1), the bias was estimated as $\bar{b_1} - \beta_1$, where $\bar{b_1}$ is the estimated log odds ratio (i.e. estimate of parameter β_1) averaged over the 1,000 simulated datasets. This was converted to percentage bias. In this analysis, because the bias was quite small in some of the scenarios, I also give the Monte Carlo standard error (MCSE) of the percent bias. The MCSE is the standard error of the estimates (in this case percent bias) across the simulations – in other words, a measure of the between simulation variability. In this analysis I also calculated the empirical standard error, the standard deviation of the point estimates for each parameter. For the IPW and MI analyses I also calculated the

percent increase in precision compared to the complete case analysis and, in addition, for the MI analyses only, the fraction of missing information (FMI).

6.2.5 Results

6.2.5.1 Complete case analysis

Tables 6-10 and 6-13, and Figures 6-1 to 6-3 show the percent bias in the log odds ratios for all the scenarios considered. There was no evidence for bias in the complete case estimate of the log odds ratio when there was a weak association between the outcome and missingness (OR for being observed for 1 SD increase in outcome = 0.9) and no interaction between the exposure and (continuous) outcome with respect to the probability of missingness (Table 6-10). When this odds ratio was 0.75, there was evidence for a small amount of bias (<2%) in the complete case estimate in some of the scenarios (Table 6-10), although at higher levels of missing data the Monte Carlo error was increased and the results consistent with there being no bias. In these scenarios there would be no bias if missingness was predicted by the binary outcome (as opposed to the underlying continuous outcome); as such this small amount of bias must be due to missingness being predicted by the underlying continuous variable.

When an interaction between the exposure and outcome with respect to the odds of missingness was introduced, the bias increased; the resulting bias ranged from an estimated 10% when 20% of the outcome data were missing up to just under 40% bias in the log odds ratio when 80% of the outcome data were missing (Table 6-13 and Figure 6-3).

6.2.5.2 Inverse probability weighting

For 20%, 40% and 60% missing data, there was no evidence for bias in the IPW estimates of the log odds ratio when there was a weak association between the outcome and missingness (OR for being observed for 1 SD increase in outcome = 0.9) and no interaction between the exposure and (continuous) outcome with respect to the probability of missingness. When there was 80% missing data, some bias was

introduced in these scenarios (3-4% bias in the log odds ratio, Table 6-10)). When the association between the outcome and missingness was stronger (odds ratio for being observed for 1 SD increase in outcome = 0.75), bias in the IPW estimates using the proxy with low sensitivity was similar to the bias in the complete case estimates with up to 60% missing data. Using a proxy with higher sensitivity lead to increased bias in the IPW estimates (up to 8% with 80% missing data). In all scenarios (without the interaction) the standard errors of the IPW estimates were the same or very close to the standard errors of the complete case estimates. These results are shown in Table 6-10.

When the interaction was introduced between the exposure and outcome, the IPW results (when the interaction between the exposure and linked depression variable was included in the model used to generate the weights) were slightly less biased than the complete case estimates when the proxy had low sensitivity (25%). When the sensitivity was high (90%) the reductions in bias were greater. Precision was also increased. When the interaction was excluded from the model used to generate the weights, the IPW estimates were all slightly more biased than the complete case estimates (Table 6-13).

6.2.5.3 Multiple imputation

6.2.5.3.1 Passive imputation of binary outcome

When there was no interaction between the outcome and exposure with respect to the odds of missingness and the binary outcome was imputed passively (by dichotomising the imputed continuous outcome) bias was increased relative to the estimates from the complete case analysis (which, as described above, were not biased or biased by less than 2%). This bias increased as the linked proxy became stronger – i.e. as the sensitivity and specificity increased; precision was also increased (relative to the complete case analysis) with a stronger proxy. Bias and gains in precision also increased as the percentage of missing data increased (Table 6-11, Table 6-12, Figure 6-1 and Figure 6-2).

When an interaction was introduced between the exposure and outcome with respect to the odds of being observed, the MI estimates were slightly less biased than those obtained in the complete case analysis (which in this case were biased, as expected) and, in these scenarios, a proxy with 90% sensitivity resulted in slightly lower bias compared to one with 25% sensitivity; there was little difference between 95% and 99% specificity in terms of the resulting bias. Although bias was only slightly reduced compared to the complete case analysis, the estimates of the log odds ratio were estimated much more precisely, particularly at high levels of missing data (Table 6-14 and Figure 6-3).

6.2.5.3.2 Direct imputation of binary outcome

In scenarios with no interaction between the exposure and outcome with respect to missingness, imputing the binary outcome directly resulted in very similar bias (slightly lower/higher in some scenarios) compared to the complete case analysis. In contrast to when the binary outcome was passively imputed, a stronger proxy (greater sensitivity and specificity) generally resulted in slightly lower bias than with a weaker proxy (Table 6-11, Table 6-12, Figure 6-1 and Figure 6-2). When the interaction was introduced, the results obtained when imputing the binary outcome directly were very similar to those obtained using IPW: there were small reductions in bias (compared to the complete case analysis) when the proxy had low sensitivity but greater reductions when the proxy had a high sensitivity. Precision was also increased. These results are shown in Table 6-14 and Figure 6-3.

The FMI when the binary outcome was imputed directly was higher than when it was imputed passively and, for a given percentage of missing data, was lowest when precision was increased the most (i.e. when the proxy had high sensitivity for passive imputation and high sensitivity and specificity for direct imputation) (Table 6-11, Table 6-12 and Table 6-14).

Table 6-10: Complete case and IPW estimates of the log odds ratio (true log odds ratio = 0.402)

(with no interaction between exposure and outcome with respect to missingness)

Factor 1: % missing	Complete case		Factors 3,4: sensitivity, specificity	IPW		
	Estimate (empirical SE)	% bias (mcse[1])		Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]
Factor 2: OR_{obs} for 1 SD increase in continuous outcome = 0.90						
20%	0.404 (0.094)	0.5% (0.7%)	25, 95	0.404 (0.094)	0.5 (0.7%)	0%
			90, 95	0.406 (0.094)	0.9 (0.7%)	0%
			25, 99	0.405 (0.095)	0.6 (0.7%)	0%
			90, 99	0.407 (0.094)	1.1 (0.7%)	0%
40%	0.408 (0.110)	1.5% (0.9%)	25, 95	0.403 (0.110)	0.3 (0.9%)	0%
			90, 95	0.406 (0.110)	0.9 (0.9%)	0%
			25, 99	0.404 (0.110)	0.5 (0.9%)	0%
			90, 99	0.407 (0.110)	1.2 (0.9%)	0%
60%	0.407 (0.140)	1.2% (1.1%)	25, 95	0.402 (0.140)	-0.2 (1.1%)	0%
			90, 95	0.405 (0.140)	0.6 (1.1%)	0%
			25, 99	0.403 (0.141)	0.1 (1.1%)	-1%
			90, 99	0.406 (0.140)	0.9 (1.1%)	0%
80%	0.396 (0.203)	-1.6% (1.6%)	25, 95	0.414 (0.204)	2.9 (1.6%)	-1%
			90, 95	0.417 (0.203)	3.5 (1.6%)	0%
			25, 99	0.415 (0.204)	3.2 (1.6%)	-1%
			90, 99	0.418 (0.203)	4.0 (1.6%)	0%
Factor 2: OR_{obs} for 1 SD increase in continuous outcome = 0.75						
20%	0.410 (0.096)	1.9% (0.8%)	25, 95	0.407 (0.097)	1.0 (0.8%)	0%
			90, 95	0.412 (0.097)	2.4 (0.8%)	0%
			25, 99	0.408 (0.097)	1.6 (0.8%)	0%
			90, 99	0.414 (0.097)	2.1 (0.8%)	0%
40%	0.411 (0.116)	2.1% (0.9%)	25, 95	0.408 (0.115)	1.5 (0.9%)	0%
			90, 95	0.417 (0.115)	3.6 (0.9%)	0%
			25, 99	0.411 (0.115)	2.2 (0.9%	0%
			90, 99	0.420 (0.115)	4.5 (0.9%)	0%
60%	0.409 (0.149)	1.8% (1.2%)	25, 95	0.409 (0.153)	1.7 (1.2%)	-2%
			90, 95	0.418 (0.153)	3.9 (1.2%)	-2%
			25, 99	0.413 (0.153)	2.6 (1.2%)	-2%
			90, 99	0.422 (0.153)	5.0 (1.2%)	-2%
80%	0.400 (0.224)	-0.6% (1.8%)	25, 95	0.422 (0.229)	4.9 (1.8%)	-2%
			90, 95	0.429 (0.227)	6.5 (1.8%)	-1%
			25, 99	0.426 (0.231)	5.9 (1.8%)	-4%
			90, 99	0.434 (0.227)	7.8 (1.8%)	-1%

1. Monte Carlo standard error
2. Relative to complete case analysis

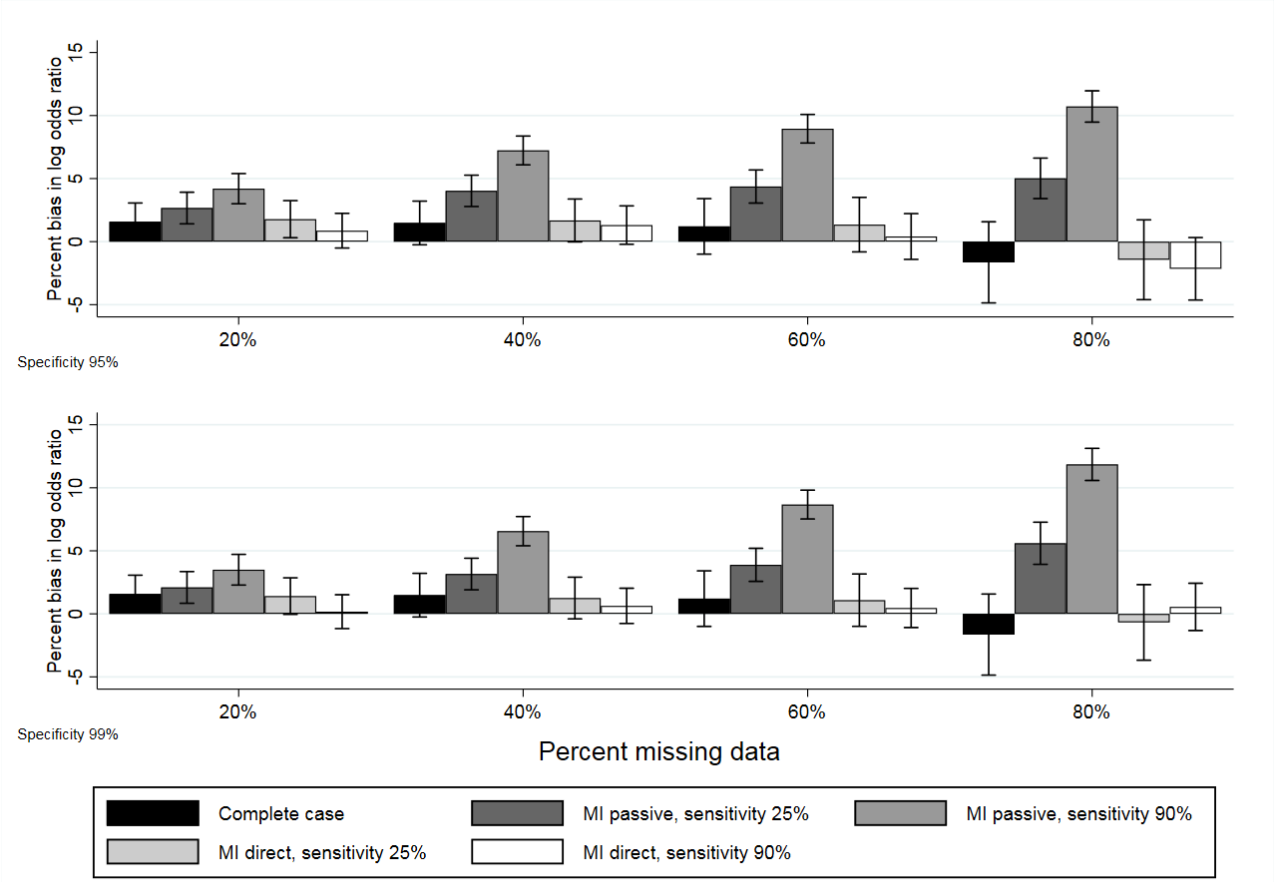


Figure 6-1: Percent bias in log odds ratio when OR_{obs} for 1 SD increase in outcome = 0.90: complete case and MI estimates

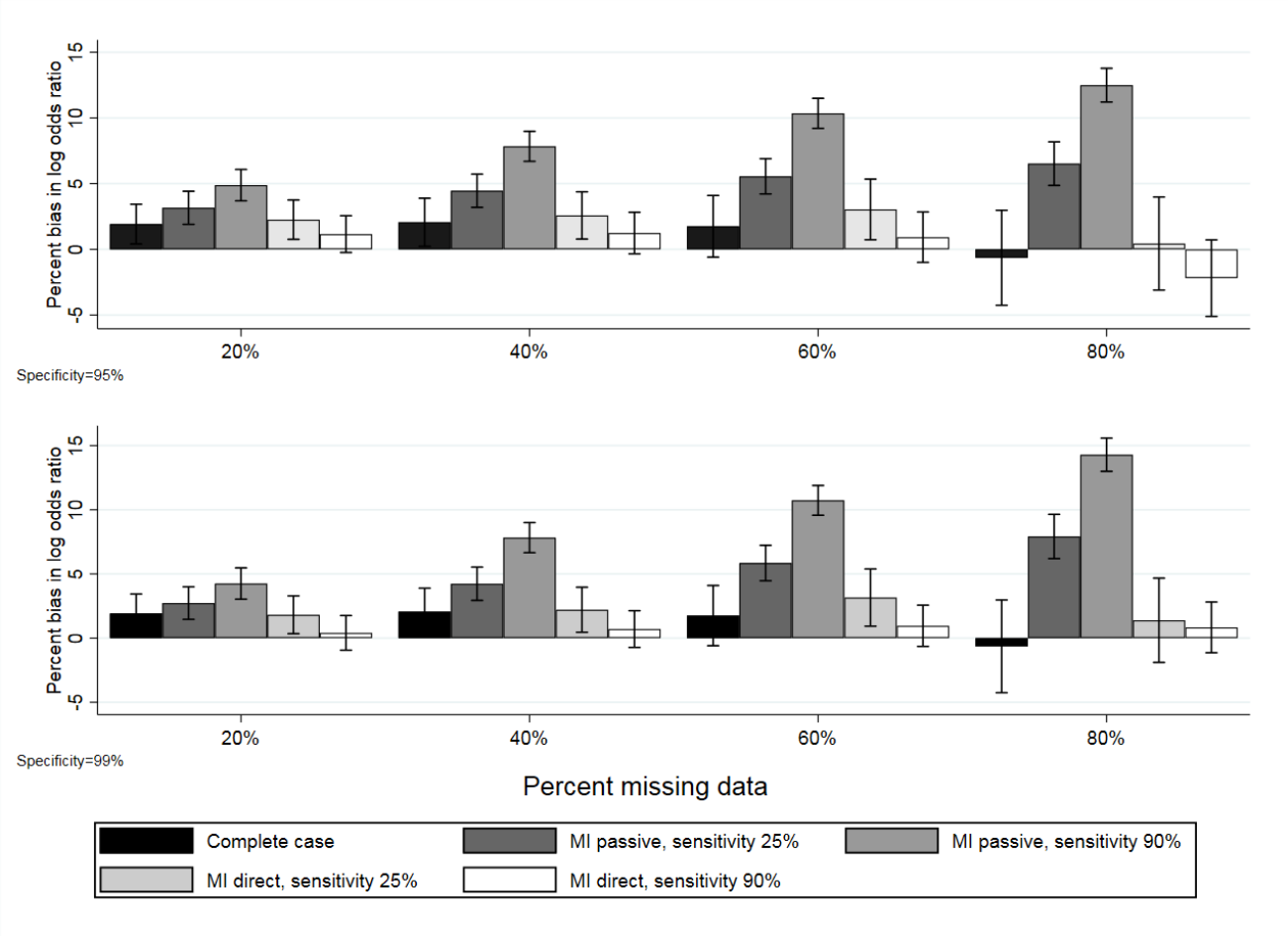


Figure 6-2: Percent bias in log odds ratio when OR_{obs} for 1 SD increase in outcome = 0.75: complete case and MI estimates

Table 6-11: MI estimates of the log odds ratio when OR_{obs} for 1 SD increase in outcome = 0.90 (Factor 2) (true log odds ratio = 0.402) (and no interaction between exposure and outcome with respect to missingness)

Factor 1: % missing	Factors 3,4: sensitivity, specificity	Binary outcome imputed passively				Binary outcome imputed directly			
		Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]	Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]
20%	25, 95	0.413 (0.080)	2.7% (0.6%)	40%	16%	0.409 (0.093)	1.8% (0.7%)	2%	21%
	90, 95	0.419 (0.076)	4.2% (0.6%)	53%	13%	0.406 (0.087)	0.9% (0.7%)	17%	12%
	25, 99	0.411 (0.080)	2.1% (0.6%)	39%	17%	0.408 (0.092)	1.4% (0.7%)	5%	19%
	90, 99	0.416 (0.078)	3.5% (0.6%)	48%	12%	0.403 (0.085)	0.2% (0.7%)	22%	6%
40%	25, 95	0.419 (0.079)	4.0% (0.6%)	94%	30%	0.409 (0.108)	1.7% (0.9%)	3%	41%
	90, 95	0.431 (0.073)	7.2% (0.6%)	128%	23%	0.408 (0.097)	1.3% (0.8%)	28%	28%
	25, 99	0.415 (0.080)	3.2% (0.6%)	91%	31%	0.407 (0.105)	1.3% (0.8%)	8%	38%
	90, 99	0.429 (0.074)	6.6% (0.6%)	123%	23%	0.405 (0.089)	0.6% (0.7%)	52%	14%
60%	25, 95	0.420 (0.084)	4.4% (0.7%)	181%	42%	0.408 (0.137)	1.3% (1.1%)	4%	61%
	90, 95	0.438 (0.072)	9.0% (0.6%)	278%	33%	0.404 (0.115)	0.4% (0.9%)	47%	46%
	25, 99	0.418 (0.085)	3.9% (0.7%)	174%	44%	0.407 (0.132)	1.1% (1.0%)	12%	58%
	90, 99	0.437 (0.073)	8.7% (0.6%)	271%	34%	0.404 (0.099)	0.5% (0.8%)	101%	27%
80%	25, 95	0.423 (0.102)	5.0% (0.8%)	300%	58%	0.397 (0.202)	-1.4% (1.6%)	3%	81%
	90, 95	0.445 (0.079)	10.7% (0.6%)	570%	47%	0.394 (0.158)	-2.2% (1.2%)	68%	71%
	25, 99	0.425 (0.106)	5.6% (0.8%)	270%	61%	0.400 (0.191)	-0.7% (1.5%)	15%	79%
	90, 99	0.450 (0.081)	11.9% (0.6%)	538%	48%	0.405 (0.119)	0.5% (0.9%)	194%	53%

1. Monte Carlo standard error
2. Relative to complete case analysis
3. Fraction of missing information

Table 6-12: MI estimates of the log odds ratio when OR_{obs} for 1 SD increase in outcome = 0.75 (Factor 2) (true log odds ratio = 0.402) (and no interaction between exposure and outcome with respect to missingness)

Factor 1: % missing	Factors 3,4: sensitivity, specificity	Binary outcome imputed passively				Binary outcome imputed directly			
		Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]	Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]
20%	25, 95	0.415 (0.080)	3.2 (0.6)	44%	17%	0.411 (0.095)	2.3 (0.7)	2%	22%
	90, 95	0.422 (0.076)	4.9 (0.6)	60%	14%	0.407 (0.089)	1.2 (0.7)	18%	14%
	25, 99	0.413 (0.081)	2.7 (0.6)	42%	17%	0.410 (0.093)	1.8 (0.7)	6%	19%
	90, 99	0.419 (0.077)	4.3 (0.6)	55%	13%	0.404 (0.086)	0.4 (0.7)	26%	6%
40%	25, 95	0.420 (0.080)	4.5 (0.6)	109%	30%	0.413 (0.114)	2.6 (0.9)	3%	42%
	90, 95	0.434 (0.073)	7.8 (0.6)	155%	25%	0.407 (0.100)	1.2 (0.8)	34%	31%
	25, 99	0.419 (0.082)	4.2 (0.6)	100%	31%	0.411 (0.111)	2.2 (0.9)	9%	39%
	90, 99	0.434 (0.075)	7.8 (0.6)	145%	24%	0.405 (0.091)	0.7 (0.7)	64%	16%
60%	25, 95	0.425 (0.086)	5.6 (0.7)	206%	42%	0.415 (0.147)	3.0 (1.2)	4%	62%
	90, 95	0.444 (0.073)	10.3 (0.6)	322%	35%	0.406 (0.122)	0.9 (1.0)	50%	51%
	25, 99	0.426 (0.088)	5.9 (0.7)	187%	44%	0.415 (0.142)	3.2 (1.1)	11%	59%
	90, 99	0.446 (0.074)	10.7 (0.6)	305%	34%	0.406 (0.102)	1.0 (0.8)	113%	30%
80%	25, 95	0.429 (0.106)	6.5 (0.8)	373%	57%	0.404 (0.225)	0.4 (1.8)	4%	82%
	90, 95	0.453 (0.081)	12.5 (0.6)	697%	48%	0.394 (0.176)	-2.2 (1.5)	71%	75%
	25, 99	0.434 (0.110)	7.9 (0.9)	338%	59%	0.408 (0.209)	1.4 (1.6)	21%	79%
	90, 99	0.460 (0.082)	14.3 (0.6)	682%	48%	0.406 (0.125)	0.8 (1.0)	236%	56%

1. Monte Carlo standard error
2. Relative to complete case analysis
3. Fraction of missing information

Table 6-13: Complete case and IPW estimates of the log odds ratio with an interaction between the exposure and outcome with respect to missingness (Factor 5) (true log odds ratio = 0.402)

Factor 1: % missing	Complete case		Factors 3,4: sensitivity, specificity	IPW –interaction included[1]			IPW – interaction not included[1]		
	Estimate (empirical SE)	% bias (mcse[2])		Estimate (empirical SE)	% bias (mcse[2])	Gain in precis-ion[3]	Estimate (empirical SE)	% bias (mcse[2])	Gain in precis-ion[3]
20%	0.442 (0.096)	10% (0.8%)	25, 95	0.438 (0.092)	9% (0.7%)	1%	0.444 (0.096)	11% (0.7%)	0%
			90, 95	0.414 (0.086)	3% (0.7%)	15%	0.448 (0.096)	12% (0.7%)	0%
			25, 99	0.437 (0.090)	8% (0.7%)	5%	0.445 (0.096)	11% (0.7%)	0%
			90, 99	0.412 (0.081)	2% (0.6%)	29%	0.450 (0.092)	12% (0.7%)	0%
40%	0.481 (0.112)	20% (0.9%)	25, 95	0.471 (0.106)	17% (0.8%)	2%	0.485 (0.112)	21% (0.8%)	0%
			90, 95	0.425 (0.094)	6% (0.7%)	29%	0.490 (0.112)	22% (0.8%)	0%
			25, 99	0.467 (0.103)	16% (0.8%)	8%	0.486 (0.112)	21% (0.8%)	0%
			90, 99	0.418 (0.086)	4% (0.7%)	55%	0.492 (0.112)	22% (0.8%)	0%
60%	0.514 (0.140)	28% (1.1%)	25, 95	0.504 (0.135)	25% (1.1%)	3%	0.525 (0.140)	30% (1.1%)	0%
			90, 95	0.433 (0.111)	8% (0.9%)	53%	0.530 (0.140)	32% (1.1%)	0%
			25, 99	0.450 (0.130)	23% (1.0%)	12%	0.526 (0.140)	31% (1.1%)	0%
			90, 99	0.421 (0.097)	5% (0.8%)	100%	0.533 (0.140)	32% (1.1%)	0%
80%	0.553 (0.200)	38% (1.6%)	25, 95	0.538 (0.197)	34% (1.5%)	2%	0.566 (0.201)	41% (1.6%)	-1%
			90, 95	0.443 (0.152)	10% (1.2%)	69%	0.569 (0.200)	41% (1.6%)	0%
			25, 99	0.525 (0.184)	31% (1.4%)	15%	0.569 (0.200)	41% (1.6%)	0%
			90, 99	0.424 (0.117)	5% (0.9%)	190%	0.571 (0.200)	42% (1.6%)	0%

1. In logistic model used to generate the inverse probability weights
2. Monte Carlo standard error
3. Relative to complete case analysis

Table 6-14: MI estimates of the log odds ratio with an interaction between the exposure and outcome with respect to the odds of missingness (Factor 5) (true log odds ratio = 0.402)

Factor 1: % missing	Factors 3,4: sensitivity, specificity	Binary outcome imputed passively				Binary outcome imputed directly (not passive)			
		Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]	Estimate (empirical SE)	% bias (mcse[1])	Gain in precis-ion[2]	FMI[3]
20%	25, 95	0.437 (0.081)	9% (0.6%)	40%	17%	0.437 (0.095)	9% (0.7%)	1%	22%
	90, 95	0.434 (0.077)	8% (0.6%)	54%	13%	0.412 (0.089)	2% (0.7%)	15%	13%
	25, 99	0.438 (0.080)	9% (0.6%)	41%	18%	0.434 (0.094)	8% (0.7%)	3%	20%
	90, 99	0.432 (0.077)	8% (0.6%)	53%	13%	0.408 (0.086)	1% (0.7%)	24%	6%
40%	25, 95	0.471 (0.077)	17% (0.6%)	113%	31%	0.470 (0.109)	17% (0.9%)	5%	43%
	90, 95	0.460 (0.070)	14% (0.6%)	153%	24%	0.420 (0.095)	5% (0.7%)	38%	29%
	25, 99	0.475 (0.078)	18% (0.6%)	105%	32%	0.466 (0.108)	16% (0.8%)	8%	40%
	90, 99	0.461 (0.073)	15% (0.6%)	137%	24%	0.412 (0.090)	3% (0.7%)	53%	15%
60%	25, 95	0.498 (0.085)	24% (0.7%)	167%	44%	0.497 (0.138)	24% (1.1%)	2%	63%
	90, 95	0.483 (0.073)	20% (0.6%)	268%	34%	0.426 (0.117)	6% (0.9%)	43%	48%
	25, 99	0.507 (0.085)	26% (0.7%)	170%	45%	0.491 (0.131)	22% (1.0%)	13%	60%
	90, 99	0.488 (0.072)	21% (0.6%)	273%	35%	0.419 (0.099)	4% (0.8%)	97%	29%
80%	25, 95	0.532 (0.102)	32% (0.8%)	282%	60%	0.530 (0.197)	32% (1.6%)	3%	82%
	90, 95	0.508 (0.078)	26% (0.6%)	557%	49%	0.436 (0.156)	8% (1.2%)	64%	71%
	25, 99	0.544 (0.104)	35% (0.8%)	268%	62%	0.519 (0.182)	29% (1.4%)	21%	80%
	90, 99	0.517 (0.078)	29% (0.6%)	559%	50%	0.425 (0.119)	6% (0.9%)	181%	54%

1. Monte Carlo standard error
2. Relative to complete case analysis
3. Fraction of missing information

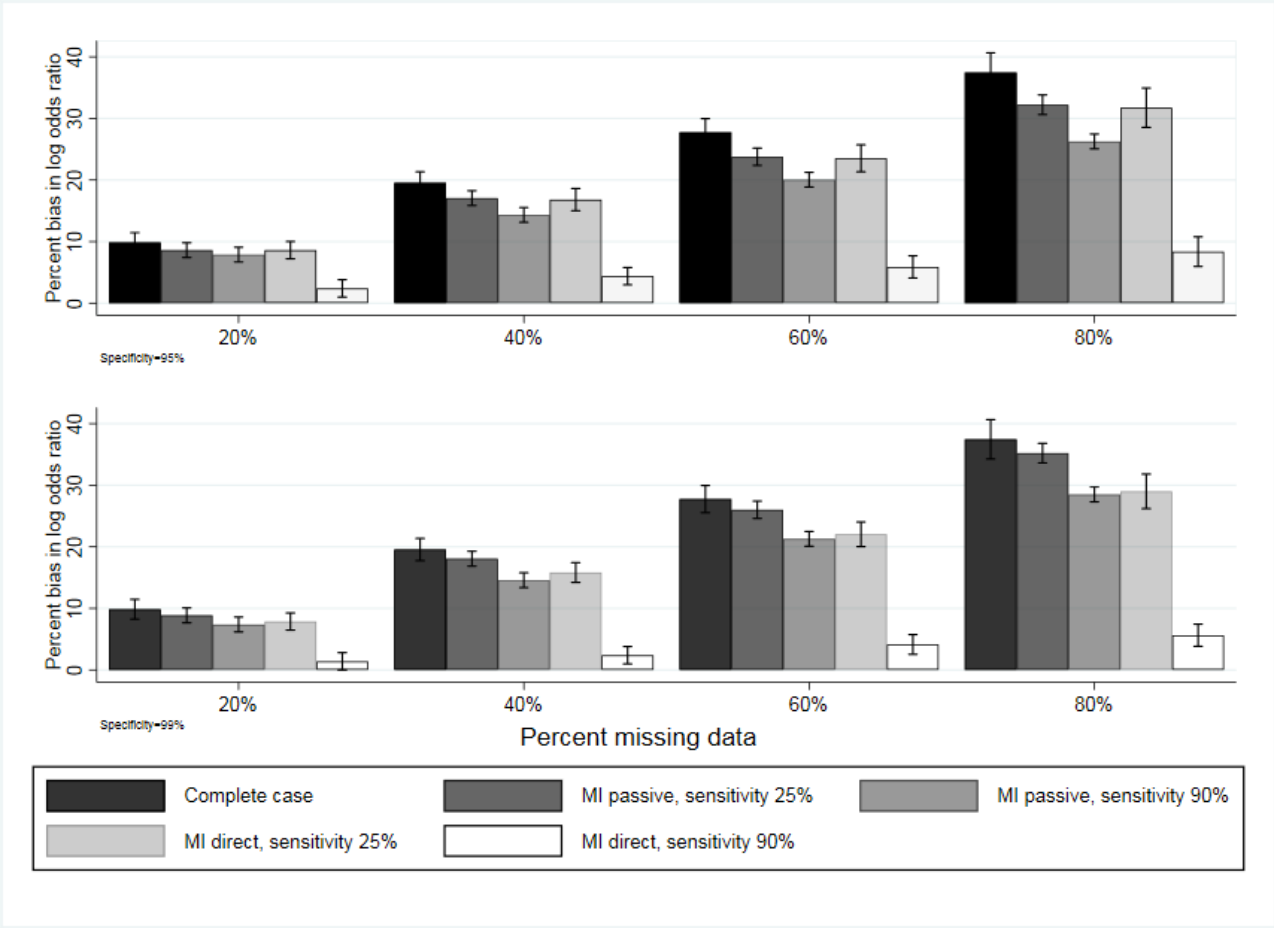


Figure 6-3: Percent bias in log odds ratio with interaction between exposure and outcome with respect to missingness: complete case and MI estimates

6.3 Discussion

The results from the simulations suggest that when missingness in a binary outcome variable is predicted by an underlying continuous measure and there is no interaction between the exposure and outcome with respect to the probability of the outcome being observed, then estimates of the log odds ratio for exposure will be subject to little or no bias; if there is an interaction, the bias could be substantial. In the absence of an interaction, having an imperfect proxy for the missing binary outcome and including it as an auxiliary variable in MI will result in similar (but more precise) estimates of the log odds ratio compared to the complete case analysis if the binary outcome is imputed directly, particularly when the proxy has high sensitivity and specificity. In contrast, if the binary outcome is imputed passively (by imputing the continuous outcome and then dichotomising the resulting imputed variable), estimates from MI are likely to be more biased than the complete case estimates (as well as being estimated much more precisely). Using IPW would not result in bias reduction or increases in precision compared to the complete case analysis.

If an interaction is present, imputing the binary outcome directly would lead to relatively large reductions in bias if the proxy had high sensitivity and specificity; otherwise the bias reductions are likely to be small. Similarly, imputing the binary outcome passively is likely to lead to only small reductions in bias. Using IPW in the presence of an interaction will lead to similar results to those obtained using MI (with direct imputation of the binary outcome) but only if an interaction term is included in the model used to generate the weights; if this is omitted, IPW could result in small increases in bias relative to the complete case analysis.

In the exemplar presented in this chapter, I used three (rather than one) linked proxies for the binary outcome; these had sensitivities of 22% (historical diagnosis or symptoms or treatment), 26% (current diagnosis or symptoms or treatment) and 55% (future diagnosis or symptoms or treatment) and specificities of 96%, 98%, and 80% (respectively). These proxies were also all independently associated with ALSPAC-

measured depression so would be better (in terms of accurately predicting ALSPAC-measured depression) than a single proxy with sensitivity of 25% but may not predict it as accurately as a single proxy with sensitivity of 90%. In the exemplar, I imputed the binary depression measure directly and, although there was no evidence for an interaction between the exposure and outcome with respect to the probability of missingness, an interaction cannot be ruled out. The MI estimate of the odds ratio in this example was slightly lower than the estimate obtained from the complete case analysis. The results of the simulations suggest that, if an interaction were present, it is likely that the MI estimate (OR=1.34; 95% CI: 0.96, 1.89) would be less biased than the complete case estimate (OR=1.47), although probably still an over-estimate. In the analysis of the ALSPAC data I did not include interaction terms (between GP measures of depression and smoking in pregnancy) in the model used to generate the weights – due to the small cell counts, this would probably have led to large reductions in precision in the IPW estimates. The IPW estimate of the odds ratio in this analysis was quite similar to the complete case estimate, as in the simulations. These results are therefore consistent with there being no impact of smoking during pregnancy on offspring depression, or at most a 34% increase in the odds of depression (under the assumption of no residual confounding).

When a continuous outcome is MNAR a complete case analysis will result in a biased estimate of the exposure-outcome association. Although I have not examined this directly, the results presented in this chapter suggest that, in contrast, if this outcome is dichotomised and a complete case logistic regression used to carry out the analysis, then this is likely to produce estimates that are subject to little or no bias (if the log odds of missingness increases linearly as the underlying continuous outcome increases and if there is no interaction between the outcome and exposure with respect to the probability of missingness). Future work could answer this specific question – with different mechanisms causing missingness.

Chapter 7 Missing categorical exposure

In this chapter I examine bias due to missing data in a categorical exposure. As before, I do this through an exemplar and a simulation study. As outlined in Chapters 1 and 2, if an exposure variable is missing then a complete case analysis will produce unbiased estimates of the exposure-outcome association as long as missingness does not depend on the outcome (Carpenter and Kenward 2013). Further, if the outcome is binary and logistic regression is used for the complete case analysis, the exposure odds ratio will be unbiased as long as there is not a multiplicative interaction between the exposure and outcome with respect to the probability of missingness (Bartlett et al. 2015). If the exposure is MNAR then multiple imputation and IPW will introduce bias (Carpenter and Kenward 2013, Seaman and White 2013).

7.1 Exemplar: teenage smoking and educational attainment

In this exemplar, described in Chapter 2 (Section 2.5.3), I am interested in whether teenage smoking has an impact on educational attainment at age 16 years. I have two outcome variables (measures of educational attainment at 16 years), one of which is numerical (an attainment score) and one of which is binary; these are listed in Table 7-1 and described in greater detail in Chapter 3 (Section 3.4.1). Thus, I have two analysis models in this exemplar: a multiple linear regression and a multiple logistic regression. I also have different measures of exposure (teenage smoking). Again, these are listed in Table 7-1. As in Chapter 6, the linked data for this exemplar come from GP records.

7.1.1 Analysis

Subjects included in this analysis were all singletons and twins enrolled in the study who were alive at one year, had not explicitly dissented to linkage to their health records, and for whom ALSPAC had a valid NHS ID number (n=14,566).

7.1.1.1 Variables

The variables included in this analysis are summarised in the table below.

Table 7-1: Variables included in the analysis of Exemplar 3

Variable	Details	Chapter 3 section
Outcomes	(1) Continuous: capped GCSE / equivalent score (KS4 attainment score) (2) Binary: whether an individual did not or did obtain five or more A* to C grades at GCSE / equivalent (1=did not, 0=did)	3.3.1
Exposures	Teenage smoking (at 15 years): Ever-smoked Frequency of current smoking Serum cotinine >9.5ng/ml vs ≤9.5ng/ml	3.3.2
Covariates	Child's sex and KS2 attainment score; maternal age (at the child's birth and at first pregnancy), parity, marital status, ethnicity and smoking; mother's and father's educational level; father's smoking (ever smoked) family occupational social class; housing tenure; use of car; phone in home; double glazing; number of rooms in the house and financial difficulties score.	3.3.3
Linked GP variables	GP record of smoking - four different variables: Ever smoked before age 16 years (any record of smoking/ex-smoking before age 16); current smoking while aged 15; smoker at age 16-19 years; smoker at age 20+ years (see Section 7.1.1.1.1 for details)	3.4.2.2

Although there was no evidence that age at first pregnancy was associated with either KS4 attainment, teenage smoking, or missingness in smoking in the analysis carried out for this chapter, it was strongly associated with missingness in the GP measures of smoking. Therefore, this variable was omitted from the complete case analysis but included in the IPW and MI models that included the linked GP variables.

As in Chapters 5 and 6, so that the MI models would converge, some of the maternal age and age at first pregnancy categories were combined. I used the new maternal age variable in the complete case analysis so that the results from this and the IPW/MI models were comparable (i.e. adjusted for identical variables).

7.1.1.1.1 Linked GP data

In Chapter 3 (Section 3.4.2.2) I described the codes used in two recent studies to define smoking status using GP data; in my analysis I used the combined list of codes from the two studies, with a few minor changes:

- I did not include the codes 6791. (health ed – smoking) or 67910 (health ed – parental smoking) as this appeared to generate a large number of false positives (for example, there were many occasions on which this was recorded alongside a code for never smoked).

- Similarly, if other codes about smoking cessation advice (67H1., 67H6., 8CAL., and others) were recorded but the individual was concurrently recorded as having never smoked, then I recorded this as not smoking.

As in the study by Atkinson and colleagues (Atkinson et al. 2017), if a Read code required a value to be recorded (e.g. number of cigarettes per day) then I only classified someone as a smoker according to this code if this value was non-missing and greater than zero. If the value was missing or zero then smoking status was coded as being uncertain, unless the individual was subsequently or concurrently recorded as having never smoked (Read code 1371.), in which case they were coded as not smoking.

Thus, at one or more time points individuals were recorded as either a non-smoker, an ex-smoker, or a smoker. From these, I created the four different smoking variables listed in Table 7-1. For each of these, smoking was recorded as positive if they were recorded as a smoker at that age (and/or ex-smoker for the first variable, ever smoked), negative if there were recorded as a non-smoker (or they had no smoking records at all) AND the individual was still present in the GP data up to and including

the upper age limit applicable for that variable (so they had to still be registered in a contributing practice up to their 16th birthday for the first of these two variables, for example).

7.1.1.2 Association between ALSPAC (self-reported) smoking and GP-recorded smoking

The association between the ALSPAC recorded smoking variables (self-reported smoking and cotinine) and GP-recorded smoking was examined using 2 x 2 tables. Since I hypothesised that these associations might be different for males and females, I looked at these associations separately by sex.

7.1.1.3 Examining the missing data mechanism

Logistic regression was used to examine the predictors of missingness. The factors included were those identified in Chapter 4; the linked smoking variables were included to investigate whether there was evidence for ALSPAC-recorded smoking being MNAR.

7.1.1.4 Dealing with missing data

Three different approaches were used to deal with missing data when modelling the relationship between smoking and attainment:

a) A complete case analysis
b) Inverse probability weighting, both including and excluding the linked smoking variables.
c) Multiple imputation using chained equations, also performed both with and without the linked variables.

7.1.1.4.1 Inverse probability weighting

I used two logistic models used to obtain the inverse probability weights; both included the baseline covariates listed above. Model 1 only included these variables; model 2 also included the linked GP variables. The four GP-derived smoking variables were obviously strongly associated with each other. In order to avoid small cell counts when predicting the inverse probability weights, I chose two of them (to

include in both the IPW models and the MI models). Smoking before age 16 gave higher sensitivities but similar specificities to smoking at age 15 and smoking aged 16-19 was more strongly associated with the ALSPAC measures of smoking than smoking aged 20 or older. Thus, these were the two GP-derived smoking variables I included. Weights were calculated separately for males and females. As previously, the Hosmer-Lemeshow goodness of fit test was used to assess the fit of the logistic models used to generate the (inverse probability) weights. Finally, as a sensitivity analysis, large weights were truncated – choosing maximum values 10, 8, 6 and 4.

7.1.1.4.2 MI models

I also used two MI models. The MI models included the same variables as included in the two IPW models (described above) – so one model with only baseline covariates and one model with the two GP measures of smoking. This was done to make the MI and IPW models comparable in terms of the auxiliary variables included. Additional auxiliary variables from ALSPAC (apart from age at first pregnancy, as described above) were not included in either the IPW or the MI models because the focus in this analysis was on the impact and utility of the linked variables as auxiliary variables in these models. All imputations were carried out separately for males and females.

7.1.2 Results

Of the 14,566 subjects included in this analysis, 5,273 (36.2%) had data on ever smoking (by age 15 years), 5,328 (36.6%) on frequency of smoking at 15 years and 3,441 (23.6%) had cotinine measured at age 15 years. Between 6,560 (45.0%) and 9,560 (65.6%) had GP data on individual smoking variables, depending on the age at which it was defined (i.e. the lower number here was for smoking at age 20 years or over; the higher number was for smoking before age 16). A total of 10,032 (68.9%) individuals had data on at least one (GP) smoking variable. Altogether, 11,973 (82.2%) individuals had non-missing data on the KS4 attainment score and 12,097 (83.0%) had non-missing data on the binary attainment variable (not obtaining five or more A* to C grades). There were 2,527 (17.4%), 2,596 (17.8%) and 1,682 (11.5%) complete cases

for the analysis of ever smoking, frequency of smoking and cotinine status at age 15 years, respectively. Further details are given in Table 7-2.

Table 7-2: Completeness of ALSPAC data by availability of linked GP data

Complete data on:			Linked GP data		Total
Covariates	Ever smoked	Outcomes	Yes[1]	No	
Yes	Yes	Yes	2,057 (81%)	470 (19%)	2,527
		No	78 (73%)	29 (27%)	107
	No	Yes	1,944 (70%)	822 (30%)	2,766
		No	88 (51%)	86 (49%)	174
No	Yes	Yes	1,666 (79%)	432 (21%)	2,098
		No	301 (56%)	240 (44%)	541
	No	Yes	3,174 (69%)	1,408 (31%)	4,582
		No	724 (41%)	1,047 (59%)	1,771
	Frequency of smoking				
Yes	Yes	Yes	2,118 (82%)	478 (18%)	2,596
		No	76 (77%)	23 (23%)	99
	No	Yes	1,883 (70%)	814 (30%)	2,697
		No	90 (49%)	92 (51%)	182
No	Yes	Yes	1,631 (79%)	442 (21%)	2,073
		No	310 (55%)	250 (45%)	560
	No	Yes	3,209 (70%)	1,398 (30%)	4,607
		No	715 (41%)	1,037 (59%)	1,752
	Cotinine				
Yes	Yes	Yes	1,363 (81%)	318 (19%)	1,681
		No	46 (57%)	35 (43%)	81
	No	Yes	2,636 (73%)	974 (27%)	3,610
		No	119 (54%)	100 (46%)	219
No	Yes	Yes	1,024 (78%)	284 (22%)	1,308
		No	224 (57%)	167 (43%)	391
	No	Yes	3,818 (71%)	1,556 (29%)	5,374
		No	802 (42%)	1,120 (58%)	1,922
Total			10,032 (79%)	4,534 (31%)	14,566

1. Information on at least one of: smoking before age 16, smoking at age 15, smoking at age 16-19, smoking at age 20 years or older

7.1.2.1 Association between ALSPAC-recorded and GP-recorded smoking

Table 7-3 shows the association between the three ALSPAC measures of smoking and GP-recorded smoking. The GP measure used in each comparison was selected to

correspond most closely (in terms of its definition) with the ALSPAC measure: so, ever smoked was compared to any GP record of smoking before age 16; the other two ALSPAC smoking variables were compared to GP smoking record while aged 15. The comparisons are shown separately for males and females. Among those classified as non-smokers in ALSPAC, at least 98% were recorded as non-smokers according to the GP data. Among females who were classified as ever smokers in ALSPAC at age 15 years, only 14% were recorded as having ever smoked in the GP data. For daily smokers, the proportion of females identified as smoking at age 15 in the GP data was 37%; similarly among those with a cotinine level >9.5ng/ml, 31% were identified as smoking at age 15. The corresponding figures among males were lower: 10% of ever smokers, 22% of daily smokers and 16% of those with cotinine >9.5ng/ml (Table 7-3).

Table 7-3: Comparison of ALSPAC-recorded and GP-recorded smoking by sex

ALSPAC measure	Females			Males		
	GP: Any record of smoking before age 16 years					
Ever smoked?	No/ Uncertain[1]	Yes	Total	No/ Uncertain[1]	Yes	Total
No	895 (99%)	6 (1%)	901	>99%[2]	<1%[2]	884
Yes	1169 (87%)	176 (13%)	1,345	752 (91%)	70 (9%)	822
Frequency of smoking	GP: Current smoking at age 15[3]					
	No/ Uncertain[1]	Yes	Total	No/ Uncertain[1]	Yes	Total
Never	1,705 (99%)	20 (1%)	1,725	>99%[2]	<1%[2]	1,396
< Daily	252 (94%)	15 (6%)	267	149 (97%)	5 (3%)	154
Daily	120 (67%)	59 (33%)	179	74 (78%)	21 (22%)	95
Cotinine	GP: Current smoking at age 15[3]					
	No / Uncertain[1]	Yes	Total	No/ Uncertain[1]	Yes	Total
≤9.5ng/ml	1,181 (98%)	21 (2%)	1,202	>99%[2]	<1%[2]	1,137
>9.5ng/ml	95 (73%)	36 (27%)	131	87 (84%)	16 (16%)	103

1. Uncertain smoking status – smoking cessation codes but not positive record of being a current smoker OR smoking recorded with missing value for amount smoked
2. Exact numbers suppressed for disclosure control reasons
3. Any smoking before age 16 years gave similar specificities but slightly higher sensitivities (41% for female daily smokers, 24% for daily male smokers; 35% for female cotinine >9.5ng/ml, 17% for male cotinine >9.5ng/ml)

Table 7-4 show the comparisons for each of the three ALSPAC measures with smoking recorded at age 16-19 years and at 20 years or older. These measures picked up a greater proportion of smokers but, as expected, resulted in a larger number of false positives. For ever-smoking and frequency of smoking the specificity was similar for males and females, but for the cotinine measure there was a greater proportion of false negatives among females than males. For all measures, but particularly frequency of smoking and cotinine, the GP measures picked up a larger proportion of smokers among females compared to males.

Table 7-4: Comparison of ALSPAC-recorded and GP-recorded smoking by sex: later GP measures

ALSPAC measure		Females		Males	
		GP: Smoking record aged 16-19 years			
		No/ Uncertain[1]	Yes	No/ Uncertain[1]	Yes
Ever smoked?	No Yes	694 (90%) 662 (52%)	80 (10%) 608 (45%)	663 (91%) 474 (60%)	65 (9%) 315 (40%)
Frequency of smoking	Never < Daily Daily	1,278 (84%) 131 (52%) 27 (15%)	241 (16%) 123 (48%) 158 (85%)	1,176 (87%) 87 (58%) 36(34%)	182 (13%) 62 (42%) 71 (66%)
Cotinine	≤9.5ng/ml >9.5ng/ml	838 (79%) 26 (9%)	223 (21%) 109 (81%)	822 (84%) 44 (40%)	153 (16%) 65 (60%)
		GP: Smoking record aged 20 years or older			
		No/ Uncertain[1]	Yes	No/ Uncertain[1]	Yes
Ever smoked?	No Yes	274 (73%) 311 (31%)	100 (27%) 701 (69%)	339 (75%) 297 (44%)	112 (25%) 374 (56%)
Frequency of smoking	Never < Daily Daily	529 (62%) 60 (30%) 12 (8%)	320 (38%) 142 (70%) 160 (92%)	640 (70%) 55 (43%) 29 (28%)	269 (30%) 73 (57%) 73 (72%)
Cotinine	≤9.5ng/ml >9.5ng/ml	331 (54%) 11 (9%)	283 (46%) 117 (91%)	444 (68%) 35 (35%)	208 (32%) 66 (65%)

1. Uncertain smoking status – smoking cessation codes but not positive record of being a current smoker OR smoking recorded with missing value for amount smoked

7.1.2.2 Predictors of missing self-reported smoking

Table 7-5 and Table 7-6 show the associations between the covariates and the outcomes and missingness in the ALSPAC smoking variables among those with

complete data on covariates and each education outcome (n=5,293 for the continuous outcome and n=5,313 for the binary outcome). Some of the covariates (maternal smoking and parity) were more strongly associated with missingness in self-reported frequency of smoking at 15 than ever-smoked at 15 (derived from smoking variables at 12,13,14 and 15) and cotinine at 15 years, whereas others (for example, breastfeeding duration) were consistently associated with each of the three smoking variables. In contrast, sex and Key Stage 2 attainment, which were not associated with missingness in cotinine after adjustment for the other covariates and the continuous outcome, were strongly associated with missingness in frequency of smoking and ever having smoked. Both outcomes (capped attainment score and 5+ A*- C grades) were associated with missingness in all three exposure variables, but these relationships were much stronger for frequency of smoking and cotinine than for ever having smoked (Table 7-5).

Table 7-5: Predictors of missingness in ALSPAC smoking: child and maternal covariates plus outcome variables

(n=5,293/n=5,313 with complete covariates and continuous/binary outcome)

Factor[1]	Level	OR (95% CI)[2]		
		Missing "ever-smoked"	Missing "frequency of smoking"	Missing cotinine
Sex	Female vs male	0.62 (0.55, 0.70)	0.78 (0.69, 0.87)	1.04 (0.92, 1.18)
Mother's education	O level/lower	1.00	1.00	1.00
	A level	0.82 (0.71, 0.95)	0.83 (0.72, 0.96)	0.80 (0.69, 0.93)
	Degree/higher	0.89 (0.72, 1.11)	0.99 (0.80, 1.24)	0.95 (0.76, 1.18)
Duration of breastfeeding	Never/<1 month	1.00	1.00	1.00
	1 to <3 months	0.61 (0.50, 0.73)	0.69 (0.57, 0.84)	0.71 (0.58, 0.87)
	3 to <6 months	0.72 (0.61, 0.85)	0.71 (0.60, 0.85)	0.57 (0.48, 0.68)
	6 months +	0.62 (0.54, 0.72)	0.60 (0.51, 0.69)	0.55 (0.47, 0.65)
Mother's age at birth	<20	1.00	1.00	1.00
	20-24	0.77 (0.46, 1.27)	0.55 (0.31, 0.97)	0.97 (0.52, 1.81)
	25-29	0.68 (0.41, 1.13)	0.45 (0.25, 0.79)	0.82 (0.44, 1.52)
	30-34	0.55 (0.33, 0.91)	0.35 (0.19, 0.63)	0.77 (0.41, 1.43)
	35+	0.58 (0.34, 0.99)	0.35 (0.19, 0.63)	0.82 (0.45, 1.56)
Maternal smoking	Never	1.00	1.00	1.00
	Yes; not in preg.	1.14 (1.00, 1.31)	1.31 (1.14, 1.50)	1.07 (0.92, 1.23)
	In pregnancy	0.97 (0.81, 1.15)	1.39 (1.09, 1.56)	1.02 (0.84, 1.23)
Parity	0	1.00	1.00	1.00
	1	1.11 (0.98, 1.27)	1.32 (1.16, 1.52)	0.99 (0.86, 1.14)
	2+	1.29 (1.08, 1.54)	1.71 (1.42, 2.05)	1.18 (0.97, 1.42)
KS2 attainment	Per 10 pt increase	0.96 (0.94, 0.98)	0.98 (0.96, 1.00)	1.00 (0.98, 1.03)
KS4 attainment[3]	Per 10 pt increase	0.99 (0.98, 1.00)	0.95 (0.94, 0.96)	0.95 (0.94, 0.96)
5+ A*- C grades[4]	No vs yes	1.19 (1.01, 1.40)	1.58 (1.34, 1.87)	1.65 (1.37, 1.99)

1. Odds ratios for maternal depression, ethnicity and marital status not shown: not found to be associated with missingness in any of the exposure variables (odds ratios all close to unity)
2. Mutually adjusted for all covariates plus the continuous outcome (capped KS4 attainment score)
3. Not adjusted for the binary outcome
4. Not adjusted for the continuous outcome

Table 7-6: Predictors of missingness in ALSPAC smoking: family and paternal covariates (n=5,293/n=5,313 with complete covariates and continuous/binary outcome)

Factor[1]	Level	OR (95% CI)[2]		
		Missing "ever-smoked"	Missing "frequency of smoking"	Missing "cotinine"
Father's education	O level or lower	1.00	1.00	1.00
	A level	1.08 (0.94, 1.23)	1.05 (0.91, 1.20)	1.05 (0.91, 1.22)
	Degree or higher	1.05 (0.87, 1.28)	1.04 (0.85, 1.27)	0.95 (0.78, 1.17)
Paternal smoking	Yes vs no	1.02 (0.85, 1.23)	1.08 (0.89, 1.30)	1.03 (0.84, 1.26)
Housing tenure	Mortgaged/owned	1.00	1.00	1.00
	Private rented	1.36 (1.02, 1.83)	1.30 (0.96, 1.76)	1.24 (0.89, 1.73)
	Council/HA/other	1.03 (0.82, 1.28)	0.96 (0.76, 1.21)	0.96 (0.75, 1.24)
Number of rooms	Per 1 room increase	0.97 (0.92, 1.02)	0.95 (0.90, 1.01)	1.02 (0.96, 1.08)
Phone in home	No vs yes	0.67 (0.50, 0.91)	0.77 (0.56, 1.05)	0.71 (0.49, 1.02)
Car use	No vs yes	0.88 (0.65, 1.19)	0.78 (0.56, 1.08)	0.78 (0.54, 1.13)
Double glazing	No vs full/partial	0.94 (0.83, 1.05)	0.92 (0.82, 1.04)	0.92 (0.81, 1.05)

1. Odds ratios for occupational social class and financial difficulties not shown: not found to be associated with missingness in any of the exposure variables (odds ratios all close to unity)
2. Mutually adjusted for all covariates and the continuous outcome (capped KS4 attainment score)

Table 7-7 shows the association between the GP smoking variables and missing ALSPAC-recorded smoking. Current GP-measured smoking at age 15 was the strongest predictor of missingness in cotinine and (ALSPAC) frequency of smoking, with those recorded as smokers in their GP data being less likely to have cotinine measured / have reported their current frequency of smoking (i.e. more likely to be missing this information). In contrast, individuals who were recorded (in the GP data) as smokers aged 16-19 and aged 20+ years were less likely to have missing data on ever-smoking (self-reported in ALSPAC).

Table 7-7: Predictors of missingness in ALSPAC smoking: GP-defined smoking variables (n=4,019, n=4,007, n=3,613, and n=2,534 with complete covariate information plus the individual linked variables, respectively)

	OR (95% CI)[1] (yes vs no/uncertain)		
GP smoking variable	Missing "ever smoked" (age 15 years)	Missing "frequency of smoking" (age 15 years)	Missing cotinine (age 15 years)
Smoking before age 16	0.85 (0.62, 1.17)	1.76 (1.26, 2.46)	1.35 (0.94, 1.94)
Current smoking at age 15	0.81 (0.57, 1.15)	1.82 (1.25, 2.66)	1.48 (0.97, 2.25)
Smoking aged 16-19	0.61 (0.52, 0.72)	1.00 (0.84, 1.18)	0.91 (0.76, 1.09)
Smoking aged 20+	0.70 (0.59, 0.84)	1.06 (0.89, 1.25)	0.94 (0.78, 1.13)

1. Adjusted for all covariates but not outcomes and not mutually adjusted

Finally, I looked at whether there was evidence for an interaction between the GP-recorded smoking variables and the binary outcome (not obtaining five or more A*-C grades) with respect to the probability of missingness. Table 7-8 shows the percentage with missing exposure data according to GP recorded smoking separately for those who obtained five or more A*- C grades and those who did not. The results are summarised below for the separate exposure variables.

Frequency of smoking

The difference in the percentage with missing data on frequency of smoking comparing (GP-recorded) smokers to non-smokers was approximately two and a half times greater among those who obtained five or more A* to C grades compared to among those who did not. After adjusting for all the baseline covariates, the odds ratio for the interaction between current GP-recorded smoking and the binary outcome (not obtaining five or more A*-C grades) was 0.75, 95% CI: 0.34, 1.66, p=0.5); the OR (for the interaction) was similar for GP-recorded smoking aged 16-19: 0.79, 95% CI: 0.56, 1.12, p=0.2.

Ever-smoking

For ever-smoking, the odds ratio for the interaction between smoking before 16 and the outcome (after adjusting for covariates) was 0.69, 95% CI 0.35, 1.34, p=0.3. There was no evidence for an interaction with smoking aged 16-19 (OR=1.04, p=0.8).

Cotinine

For cotinine there was no evidence for an interaction with either of the GP measures of smoking (OR=0.91, p=0.8 for current smoking and OR=1.01, p>0.9 for smoking aged 16-19).

These interaction tests were obtained from a logistic regression rather than a binomial regression (because the binomial regression models would not converge). Having no interaction on the logistic scale does not mean there is no interaction on the probability scale; however, I would expect the conclusions would have been similar had a binomial model been possible.

Table 7-8: Percentage with missing exposure data by GP-recorded smoking and KS4 attainment

Exposure variable (variable with missing data)	Obtained 5 or more A* to C grades			
	No		Yes	
	Smoker 16-19 year			
	No	Yes	No	Yes
Ever-smoked	74%	66%	51%	43%
Frequency of smoking	75%	77%	45%	50%
Cotinine	84%	85%	66%	66%
	Smoked before 16/current smoker at 15[1]			
	No	Yes	No	Yes
Ever-smoked	72%	65%	49%	48%
Frequency of smoking	75%	81%	44%	59%
Cotinine	84%	89%	65%	77%

1. Smoked before 16 for ever-smoked / current smoker at 15 for frequency of smoking and cotinine

7.1.2.3 Predictors of missing GP-recorded smoking

To have non-missing GP data on smoking, an individual had to have a GP record up to and including the age at which the variable was defined. Thus, the factors associated with missing GP-recorded smoking are not identical to those associated with having

any linked GP data, although there were several factors in common. Children whose fathers were more highly educated were more likely to have missing linked GP data on smoking as were children who were breastfed for longer. There was also some evidence that females were less likely to have missing GP data. Missingness in the GP data on smoking was also weakly associated with paternal smoking (Table 7-9). After adjusting for these factors, there was no evidence that the other factors were associated with missing linked GP data (ORs all relatively close to unity; results not shown).

Table 7-9: Predictors of missingness in linked GP smoking data (n=5,574 with complete covariates)

Factor	Level	OR (95% CI)[1]
Father's education	O level / lower	1.00
	A level	1.13 (0.97, 1.31)
	Degree / higher	1.31 (1.07, 1.61)
Mother's age at first pregnancy	<20	1.00
	20-24	1.26 (1.01, 1.59)
	25-29	1.32 (1.03, 1.69)
	30+	1.32 (0.96, 1.81)
Duration of breastfeeding	Never/<1 month	1.00
	1 to <3 months	0.94 (0.76, 1.17)
	3 to <6 months	1.14 (0.94, 1.37)
	6+ months	1.23 (1.05, 1.44)
Paternal smoking	Yes vs no	0.86 (0.70, 1.05)
Sex	Female vs male	0.89 (0.79, 1.01)

1. Mutually adjusted for all covariates (other factors not shown)

7.1.2.4 Relationship between smoking and attainment

Table 7-10 shows the relationship between teenage smoking and the KS4 attainment score and Table 7-11 the relationship with not obtaining five or more A* to C grades using the different analysis approaches. The Hosmer-Lemeshow goodness of fit tests indicated that the logistic models used to predict the inverse probability weights did not fit poorly for cotinine or frequency of smoking (for cotinine: χ^2_8 = 7.8, p=0.4

without linked variables and χ_8^2 = 4.3, p=0.8 with linked smoking variables; for frequency of smoking: χ_8^2 = 9.5, p=0.3 without linked variables and χ_8^2 = 4.4, p=0.8 with linked variables). For ever-smoking there was evidence for a poor fit when linked variables were excluded (χ_8^2 = 18.3, p=0.02) but not when the linked variables were included (χ_8^2 = 5.7, p=0.7).

7.1.2.4.1 Attainment score (continuous outcome)

For ever-smoking the results for the continuous outcome were all quite similar (Table 7-10). For frequency of smoking the complete case analysis and multiple imputation gave very similar (fully adjusted) point estimates and confidence limits, but the results from IPW were more extreme and had wider confidence intervals. However, after truncating the weights, the IPW estimates became closer to those obtained from the complete case analysis and MI (for example, when weights were truncated at 4, the adjusted estimate for daily smoking vs never/<daily was -60 (95% CI: -71, -48); full results for truncated weights given in Appendix B, Table 13). Finally, for cotinine, the adjusted estimate from the complete case analysis was closer to unity than the IPW and MI estimates. The IPW estimate was again slightly more extreme than the estimate from MI but, as before, this became more similar after truncating the weights (for example, -53 when weights were truncated at 6, -50 when truncated at 4: Appendix B, Table 13). For all exposure variables, the results from the MI models using the linked variables were very similar to those from the MI models with no linked variables.

7.1.2.4.2 Not obtaining five or more A*-C grades (binary outcome)

For the binary outcome, there were greater differences between the results. For all exposures, the adjusted odds ratio estimates from the complete case analysis were larger than those obtained from MI (both with or without the linked variables). The estimates from the IPW models were also higher than those obtained using MI. These differences were greatest for frequency of smoking, where the adjusted odds ratio comparing daily smokers to non-smokers/< daily smokers was 8.34 (5.25, 13.25) in the complete case analysis, 8.31 (4.99, 13.86) when using IPW including linked

variables, and 5.22 (4.01, 6.80) when using MI with linked variables. The MI estimates were also much more precise (for all three exposures) than those obtained from the complete case analysis and IPW models (Table 7-11). The IPW estimates became slightly smaller when large weights were truncated (Appendix B, Table 13), but the adjusted estimate for daily smoking with the weights truncated at 4 was still much larger (8.20) than the MI estimate.

Table 7-10: Mean difference in KS4 attainment score comparing exposed to unexposed individuals (for the three teenage smoking variables)

		Analysis approach				
Smoking variable (at 15 years)		Complete case (n=2,527)	Excluding linked variables		Including linked variables	
Ever-smoked			IPW[2] (n=2,527)	MI[2] (n=14,566)	IPW[3] (n=1,714)	MI[3] (n=14,566)
Unadjusted	No	0	0	0	0	0
	Yes	-50 (-55, -45)	-58 (-65, -51)	-72 (-77, -67)	-55 (-63, -47)	-72 (-77, -67)
Adjusted[1]	No	0	0	0	0	0
	Yes	-26 (-30, -23)	-29 (-33, -25)	-31 (-35, -27)	-29 (-34, -24)	-31 (-34, -27)
Frequency of smoking		(n=2,596)	(n=2,596)	(n=14,566)	(n=1,762)	(n=14,566)
Unadjusted	Never/<daily	0	0	0	0	0
	Daily	-80 (-90, -70)	-103 (-128, -78)	-102 (-113, -90)	-105 (-130, -81)	-102 (-113, -92)
Adjusted[1]	Never/<daily	0	0	0	0	0
	Daily	-51 (-57, -44)	-64 (-82, -46)	-58 (-68, -49)	-69 (-85, -53)	-57 (-66, -49)
Cotinine level		(n=1,682)	(n=1,682)	(n=14,566)	(n=1,118)	(n=14,566)
Unadjusted	<9.5 ng/ml	0	0	0	0	0
	≥9.5 ng/ml	-73 (-88, -62)	-96 (-123, -70)	-97 (-109, -86)	-92 (-116, -68)	-99 (-107, -90)
Adjusted[1]	<9.5 ng/ml	0	0	0	0	0
	≥9.5 ng/ml	-42 (-49, -35)	-55 (-72, -37)	-53 (-63, -43)	-57 (-72, -41)	-52 (-60, -44)

1. Adjusted for sex, mother's ethnicity, maternal and paternal smoking, maternal and paternal education, maternal age, parity, family occupational social class, housing tenure, duration of breastfeeding, marital status, car use, phone use, double glazing, number of rooms in home, financial difficulties score, and Key Stage 2 attainment score
2. Age at first pregnancy as an auxiliary variable
3. With the following linked variables as auxiliaries: smoking before age 16 and smoking aged 16-19 years

Table 7-11: Odds ratios for not obtaining five or more A*- C grades comparing exposed to unexposed individuals (for the three teenage smoking variables)

		Analysis approach				
Smoking variable (at 15 years)		Complete case (n=2,527)	Excluding linked variables		Including linked variables	
Ever-smoked			IPW[2] (n=2,527)	MI[2] (n=14,566)	IPW[3] (n=1,714)	MI[3] (n=14,566)
Unadjusted	No	1.00	1.00	1.00	1.00	1.00
	Yes	3.81 (3.07, 4.72)	3.83 (3.06, 4.80)	3.88 (3.35, 4.49)	3.27 (2.52, 4.23)	3.72 (3.19, 4.32)
Adjusted[1]	No	1.00	1.00	1.00	1.00	1.00
	Yes	3.32 (2.46, 4.48)	3.37 (2.49, 4.55)	2.64 (2.11, 3.31)	2.74 (1.95, 3.86)	2.56 (2.08, 3.16)
Frequency of smoking		(n=2,596)	(n=2,596)	(n=14,566)	(n=1,762)	(n=14,566)
Unadjusted	Never/<daily	1.00	1.00	1.00	1.00	1.00
	Daily	6.90 (4.94, 9.64)	7.58 (5.20, 11.04)	6.24 (4.97, 7.84)	7.74 (5.14, 11.65)	6.08 (4.94, 7.47)
Adjusted[1]	Never/<daily	1.00	1.00	1.00	1.00	1.00
	Daily	8.34 (5.25, 13.25)	7.78 (4.49, 13.50)	5.79 (4.15, 8.59)	8.31 (4.99, 13.86)	5.22 (4.01, 6.80)
Cotinine level		(n=1,682)	(n=1,682)	(n=14,566)	(n=1,118)	(n=14,566)
Unadjusted	<9.5 ng/ml	1.00	1.00	1.00	1.00	1.00
	≥9.5 ng/ml	5.40 (3.74, 7.80)	5.74 (3.84, 8.56)	5.42 (4.37, 6.72)	5.37 (3.43, 8.39)	5.54 (4.61, 6.65)
Adjusted[1]	<9.5 ng/ml	1.00	1.00	1.00	1.00	1.00
	≥9.5 ng/ml	4.88 (2.86, 8.31)	4.68 (2.64, 8.32)	4.27 (3.33, 5.49)	5.50 (3.03, 10.01)	4.11 (3.01, 5.62)

1. Adjusted for sex, mother's ethnicity, maternal and paternal smoking, maternal and paternal education, maternal age, parity, family occupational social class, housing tenure, duration of breastfeeding, marital status, car use, phone use, double glazing, number of rooms in home, financial difficulties score, and Key Stage 2 attainment score
2. Age at first pregnancy as an auxiliary variable
3. With the following linked variables as additional auxiliaries: smoking before age 16 and smoking aged 16-19 years

7.1.3 Discussion

As discussed in Chapters 1 and 2, in a linear regression, if missingness depends on both the outcome and the exposure (given the covariates) both a complete case analysis and MI are likely to produce biased estimates of the exposure-outcome association (Carpenter and Kenward 2013, White and Carlin 2010). This also holds for a binary outcome – when logistic regression is used for the analysis – unless an additional condition holds: that there is no multiplicative interaction (on the probability scale) between the exposure and (binary) outcome with respect to missingness (Bartlett et al. 2015).

In the exemplar presented above, the linked smoking variables were used as proxies for the different exposure variables to explore factors associated with missingness. In a logistic model the odds of being a complete case did indeed depend on the outcome (both the binary and the continuous outcome) and the (proxy) exposure. As such, the complete case estimate of the effect of smoking on the continuous outcome is likely to be biased. In addition, although there was no evidence for an interaction between GP-recorded smoking and the (binary) outcome with respect to the odds of being a complete case, an interaction between the exposure(s) and outcome cannot be ruled out. If this were the case, the complete case estimate of the effect of smoking on the binary outcome would also be biased.

Inclusion of the linked variables in the IPW and MI models would reduce the dependency of missingness on the exposures themselves. In particular, in MI smoking would be imputed under a model in which missingness depended mainly on covariates (as educational attainment would be a covariate rather than an outcome when imputing smoking). As such, we might expect MI (with the linked variables as auxiliaries) to be less biased than the complete case analysis.

For the continuous outcome (attainment score) the adjusted results from the complete case analysis and MI were quite similar; the IPW estimates for cotinine and ever-smoking were also similar to those from the other two analysis (but were slightly

more extreme for frequency of smoking). For the binary outcome, the adjusted odds ratios obtained from MI including linked variables were all lower than those obtained in the complete case analysis. The IPW estimates (including linked variables) were higher than the complete case estimates for cotinine and frequency of smoking but lower for ever-smoking. For the reasons outlined above, I would expect the MI estimates to be less biased than the complete case estimates. It is thus unclear whether the fact that many of the results were quite similar is because the estimates were all biased by a similar amount.

7.2 Simulations

As in the previous two chapters, this simulation study was based on the above exemplar and was carried out to investigate the impact of missingness in a categorical exposure on the exposure-outcome relationship. For simplicity I only simulated a missing binary exposure analogous to daily smoker vs non-smoker/<daily (rather than having an exposure with more than two categories) but, as in the exemplar presented in this chapter, had both a binary and a continuous outcome (analogous to the attainment score and the binary outcome in the exemplar – not obtaining vs obtaining five or more A*- C grades). As in the previous simulation studies in Chapters 5 and 6, I varied the percentage of missing data and the extent to which the outcome was MNAR. In this chapter, in addition to this, I carried out one set of scenarios with just one linked smoking variable and then a second set with two, as well as carrying out a set in which there was an interaction between the exposure and (continuous) outcome with respect to the odds of missingness. As before, complete data were simulated first; missing data were then simulated in a separate process.

7.2.1 Simulated datasets

The following eight variables were simulated, with distributions chosen to be roughly representative of those seen in ALSPAC: offspring sex, maternal education and smoking, the exposure variable (teenage smoking) and two outcome variables (Key Stage 4 attainment score and a binary variable indicating whether an individual did

not or did obtain five or more A* to C grades at Key Stage 4: 1=did not obtain 5+ A*- C grades, 0=did). The simulated exposure was based on the frequency of smoking variable (i.e. with daily smokers in one group and non-smokers plus less than daily smokers as the reference group). I simulated two linked (smoking) variables, analogous to the two variables included in the main analysis: smoked before age 16 and smoked age 16-19. As in the previous simulations, I generated 1,000 datasets of 10,000 observations.

Sex, maternal education and maternal smoking were the three covariates. Sex was drawn from a Bernoulli distribution with probability 0.5 and mother's education from a multinomial random variable with probabilities 0.6, 0.25 and 0.15, corresponding to categories O level or lower, A level, and degree or higher (respectively). Maternal smoking (with three categories: never smoked, smoked but not in pregnancy, and smoked in pregnancy) was simulated as being conditional on maternal education such that the probability of smoking decreased as education level increased. The marginal probabilities for the three maternal smoking categories were: (0.45,0.25,0.3), (0.55,0.25,0.2) and (0.65,0.25,0.1) for O level/lower, A level, and degree/higher, respectively.

The exposure variable, teenage smoking, was simulated to give a similar prevalence of daily smokers to that seen in ALSPAC, so would be analogous to the frequency of smoking variable, with non-smoker and <daily smoker combined into a single category. This was simulated as a series of Bernoulli random variables dependent on sex, maternal education and maternal smoking; the probabilities of being a daily smoker are shown in Table 7-12.

Table 7-12: Simulated probabilities of being a daily smoker

Maternal education	Maternal smoking	Male	Female
O level/lower	Never	0.07	0.1
	Yes, no in pregnancy	0.08	0.12
	In pregnancy	0.18	0.3
A level	Never	0.03	0.08
	Yes, no in pregnancy	0.06	0.1
	In pregnancy	0.08	0.2
Degree/higher	Never	0.02	0.01
	Yes, no in pregnancy	0.04	0.07
	In pregnancy	0.08	0.14

The continuous outcome, KS4 attainment score, was simulated as a standard normal variable, dependent on sex, maternal education, maternal smoking and teenage smoking as shown in Equation 1:

$$KS4_i = \beta_0 + \beta_1 \times sex_i + \beta_2 \times mumed_{1i} + \beta_3 \times mumed_{2i} + \beta_4 \times mumsmoke_{1i} + \beta_5 \times mumsmoke_{2i} + \beta_6 \times teen\_smoke_i + \varepsilon_i \quad [1]$$

This equation was also the first analysis model. In this equation, sex is the indicator variable for sex; $mumed_1$ and $mumed_2$ for the mother having A levels and a degree level qualification or higher, respectively; $mumsmoke_1$ and $mumsmoke_2$ are the indicator variables for maternal smoking (not in pregnancy and in pregnancy, respectively); teen_smoke is the indicator variable for teenage smoking and ε is the random error, following a normal distribution with mean 0 and variance σ^2, with the latter calculated to give the KS4 attainment score a total variance of 1. The coefficients of this regression model were fixed as follows: $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\beta_3 = 0.6$, $\beta_4 = -0.06$, $\beta_5 = -0.18$ and $\beta_6 = -0.8$, representing relationships similar to those seen in ALSPAC. The binary outcome variable was derived (drawn from a series of Bernoulli distributions) from this continuous score with probabilities shown in Table 7-13. The second analysis model was a logistic model with this binary outcome and the same covariates as in Equation 1. The value of the log odds ratio (for the full dataset) in this

model was 1.19, giving a (true) odds ratio for not obtaining five or more A* to C grades comparing daily smokers to non-smokers/less than daily smokers of 3.29.

Table 7-13: Simulated probabilities of NOT obtaining five or more A* to C grades

Continuous attainment score	Probability of NOT obtaining 5+ A*-C grades
<-1	1
-1 to -0.75	0.9
>-0.75 to -0.5	0.8
>-0.5 to -0.25	0.7
>-0.25 to <0	0.4
≥0 to <0.25	0.2
≥0.25 to <0.5	0.1
≥0.5	0

The linked smoking variables were simulated to give similar sensitivities and specificities to the GP smoking variables current smoking at age 15 and smoked age 16-19. That is, the first linked variable had 99% specificity for both males and females, 30% sensitivity for females and 20% sensitivity for males; the second linked variable had 85% specificity for both males and females, 85% sensitivity for females and 65% sensitivity for males.

7.2.2 Simulating the missing data

Because the focus of this investigation was the utility of proxy data for the missing exposure, I only created missing data in the exposure variable, teenage smoking. The exposure was MNAR in all scenarios. The probabilities were generated using a logistic model, again with coefficients similar to those seen in ALSPAC:

$$\begin{aligned}\text{Logit}[\Pr(\text{Teenage smoking observed})_i] \quad [2]\\ = \alpha + \gamma_1 \times sex_i + \gamma_2 \times mumed_{1i} + \gamma_3 \times mumed_{2i}\\ +\gamma_4 \times mumsmoke_{1i} + \gamma_5 \times mumed_{2i} + \gamma_6 \times teen\_smoke_i\\ + \gamma_7 \times KS4_i + \gamma_8 \times not5plus_i + \gamma_9 \times teen\_smoke_i \times KS4_i\end{aligned}$$

(not5plus is the binary outcome). Most of the regression coefficients in this model were fixed throughout the simulation study: γ_1 = ln(0.9) (female compared to male), γ_2=ln(1.3), γ_3=ln(1.45), γ_4=ln(0.7), γ_5=ln(0.55), γ_8=ln(0.9). The coefficients that were varied were α, the constant, γ_6, the coefficient for teenage smoking, γ_7 and γ_9, the coefficients for the attainment score and the interaction between teenage smoking and the attainment score (respectively). The values of α were calculated using a trial and improvement method in order to produce particular percentages of missing data. For each observation, a Bernoulli random variable with probability given by Equation 2 was drawn to determine whether teenage smoking was missing.

7.2.3 Scenarios

Key factors influencing the extent of bias are the amount of missing data and the degree to which the outcome is MNAR. Having an interaction in the missingness model will also influence the amount of bias in the complete case analysis. The strength of association between the proxy/ies and the missing study variable will determine the extent to which bias is reduced when including these in MI or IPW models. As mentioned above, I simulated some scenarios with one linked variable, and some with two. These were the factors varied in the simulations, as detailed below.

Factor 1: The percentage of missing exposure (teenage smoking) data: 20%, 40%, 60%, 80%.

Factor 2: How strongly MNAR the exposure was: odds ratio for observing smoking comparing daily smokers to <daily/non-smokers = 0.5, 0.75, and 0.9 (i.e. γ_6 from Equation [2] = ln(0.5), ln(0.75) and ln(0.9)).

Factor 3: Whether or not there was an interaction between the exposure (teenage smoking) and the continuous outcome (KS4) with respect to missingness: in models without an interaction, γ_7 was equal to ln(1.75) and γ_9 equal to 0; in models with an interaction, γ_6 was equal to ln(0.9), γ_7 was equal to ln(2) and γ_9 equal to ln(0.8).

Factor 4: One vs two linked variables: first linked variable with high specificity (99%) and low sensitivity (20% for males, 30% for females); second with lower specificity (85%) but higher sensitivity (85% for females, 65% for males).

Thus, there were sixteen scenarios with one linked smoking variable and sixteen scenarios with two linked smoking variables; in both cases there were twelve without an interaction and four with. For each scenario, 1,000 datasets were simulated.

7.2.4 Statistical Analysis

As in the analysis of the ALSPAC data described above, I estimated the coefficients for teenage smoking using a multiple linear regression model for the attainment score (KS4) and multiple logistic regression for the binary outcome (not obtaining five or more A* to C grades). These were determined using:

a) A complete case analysis
b) Inverse probability weighting. The model for the weights included both outcomes, all the covariates specified above (sex maternal smoking, and maternal education) and the linked smoking variable(s). As in Chapters 5 and 6, when an interaction (between the exposure and continuous outcome) was included in the missingness model, the model for the weights also included an interaction between the linked smoking variable(s) and the continuous outcome. As previously, I also carried out an IPW analysis in which interaction was excluded from the model for the weights.
c) Multiple imputation. For each simulated dataset, 100 imputed datasets were created. The imputation models included all the variables included in the analysis model plus the linked smoking variable(s).

As in Chapters 5 and 6, the estimates obtained from these analyses were compared to the true parameters -0.8 and 1.19. For each of the two parameters (β_6 from Equation 1 and an equivalent parameter to represent the log odds ratio for teenage smoking), the bias was estimated as $\overline{b_6} - \beta_6$, where $\overline{b_6}$ is the estimated regression

coefficient (i.e. estimate of parameter β_6) averaged over the 1,000 simulated datasets (and similar for the log odds ratio). This was converted to percentage bias. I also calculated the empirical standard error, the standard deviation of the point estimates for each parameter. For the IPW and MI analyses, I also calculated the percent increase in precision compared to the complete case analysis and, in addition, for the MI analyses only, the fraction of missing information (FMI).

7.2.5 Results

7.2.5.1 One linked smoking variable

Tables 7-14 and 7-15 show the results for both analyses when I simulated one linked smoking variable (specificity 99%, sensitivity = 30% for females, 20% for males). As the strength of association between smoking and missingness in smoking became stronger (Factor 2: odds ratio for smoking being observed comparing daily smokers to <daily/non-smokers = 0.9, 0.75, 0.5), the bias in the complete case analysis, IPW and MI models all increased; this was the case for both the continuous outcome and the binary outcome. In all cases both the MI and the IPW estimates were less biased than those obtained from the complete case analysis. When the OR for smoking being observed was 0.5, the IPW estimates were slightly more biased than the MI estimates; otherwise they were very similar.

With 20% missing data, the MI estimates had similar standard errors to the complete case estimates. At other levels of missing data, the MI estimates were more precise than the complete case estimates; the relative increase in precision increased as the percentage of missing data increased. The IPW estimates were nearly all less precise than the complete case estimates; the loss in precision increased as the percentage of missing data increased.

7.2.5.1.1 With an interaction

When an interaction was introduced between the exposure and outcome with respect to the odds of missingness, the bias increased substantially (Table 7-18). The MI estimates were less biased than those obtained from the complete case analysis

Table 7-18). Using IPW in this scenario resulted in similar levels of bias (sometimes slightly less but sometimes more) compared to the complete case analysis if an interaction was included in the model used to generate the weights (Table 7-18). If this interaction was excluded from the model for the weights, IPW always resulted in increased bias compared to the complete case analysis (Table 7-19). The estimates using IPW were less efficient than the complete case estimates in nearly all scenarios. Further, including the interaction in the model for the weights resulted in greater losses of efficiency, particularly at higher levels of missing data.

Table 7-14: Estimates of effect of daily smoking on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with one linked smoking variable: OR_{obs} comparing daily smokers to <daily/never = 0.75

	Factor 1: % missing data	Complete case analysis		IPW using single linked proxy			MI using single linked proxy			
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision[1]	Estimate (empirical SE)	% bias	Gain in precision[1]	FMI[2]
Continuous outcome (true mean diff. =-0.8)	20%	-0.73 (0.040)	9%	-0.78 (0.045)	3%	-19%	-0.78 (0.040)	3%	0%	26%
	40%	-0.71 (0.049)	12%	-0.77 (0.060)	4%	-32%	-0.78 (0.049)	3%	4%	48%
	60%	-0.71 (0.065)	11%	-0.77 (0.090)	3%	-47%	-0.78 (0.063)	2%	9%	66%
	80%	-0.74 (0.103)	8%	-0.78 (0.150)	3%	-53%	-0.79 (0.086)	2%	42%	84%
Binary outcome (true log OR = 1.19)	20%	1.10 (0.086)	-11%	1.16 (0.086)	-4%	-1%	1.15 (0.084)	-3%	3%	21%
	40%	1.08 (0.106)	-9%	1.15 (0.109)	-5%	-5%	1.15 (0.101)	-3%	12%	43%
	60%	1.09 (0.143)	-8%	1.15 (0.147)	-3%	-6%	1.16 (0.136)	-2%	11%	64%
	80%	1.15 (0.225)	-4%	1.16 (0.239)	-2%	-11%	1.18 (0.203)	-1%	23%	85%

1. Relative to complete case analysis
2. Fraction of missing information

Table 7-15: Estimates of effect of daily smoking on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with one linked smoking variable: OR_{obs} comparing daily smokers to <daily/never = 0.9 and 0.5

Factor 2: OR_{obs} for smoking = 0.9	Factor 1: % missing data	Complete case analysis		IPW using single linked proxy			MI using single linked proxy			
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]
Continuous outcome (true mean diff. =-0.8)	20%	-0.74 (0.039)	7%	-0.79 (0.043)	1%	-18%	-0.79 (0.039)	1%	-4%	26%
	40%	-0.73 (0.048)	9%	-0.79 (0.060)	1%	-33%	-0.79 (0.047)	1%	0%	47%
	60%	-0.73 (0.065)	8%	-0.79 (0.084)	1%	-40%	-0.79 (0.059)	1%	21%	66%
	80%	-0.75 (0.098)	6%	-0.79 (0.137)	1%	-51%	-0.79 (0.082)	1%	43%	84%
Binary outcome (true log OR = 1.19)	20%	1.13 (0.084)	-5%	1.18 (0.084)	-1%	0%	1.18 (0.083)	-1%	2%	21%
	40%	1.12 (0.103)	-6%	1.18 (0.104)	-1%	-5%	1.17 (0.098)	-1%	10%	43%
	60%	1.13 (0.138)	-5%	1.17 (0.140)	-1%	-7%	1.18 (0.129)	-1%	14%	64%
	80%	1.17 (0.212)	-2%	1.18 (0.224)	-1%	-10%	1.19 (0.194)	0%	19%	84%
OR_{obs} for smoking = 0.5	% missing data	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	Estimate (empirical SE)	% bias	Gain in precis-ion[1]	FMI[2]
Continuous outcome (true mean diff. =-0.8)	20%	-0.68 (0.042)	15%	-0.74 (0.048)	7%	-23%	-0.75 (0.042)	7%	-2%	28%
	40%	-0.66 (0.056)	18%	-0.73 (0.069)	8%	-41%	-0.75 (0.053)	7%	10%	49%
	60%	-0.67 (0.079)	16%	-0.75 (0.107)	7%	-51%	-0.76 (0.071)	5%	22%	68%
	80%	-0.71 (0.121)	11%	-0.76 (0.181)	5%	-55%	-0.78 (0.102)	2%	41%	85%
Binary outcome (true log OR = 1.19)	20%	1.03 (0.090)	-14%	1.10 (0.090)	-8%	-2%	1.10 (0.087)	-8%	7%	21%
	40%	1.00 (0.117)	-16%	1.08 (0.117)	-9%	-6%	1.10 (0.109)	-8%	15%	43%
	60%	1.03 (0.167)	-13%	1.10 (0.173)	-7%	-7%	1.13 (0.155)	-5%	16%	65%
	80%	1.11 (0.269)	-7%	1.14 (0.290)	-4%	-15%	1.17 (0.233)	-2%	33%	85%

1. Relative to complete case analysis
2. Fraction of missing information

7.2.5.2 Two linked smoking variables

Table 7-16 and Table 7-17 give the results for both analyses when I simulated two linked smoking variables (first linked variable as above; second linked variable with specificity 85%, sensitivity = 85% for females, 65% for males). The IPW and MI results with two linked variables were less biased than those obtained using one linked variable. Indeed, there was very little or no bias in the MI estimate for the continuous outcome in all the scenarios. For the binary outcome, there was a small amount of bias in the MI estimate (of the log odds ratio) when the odds ratio for smoking being observed comparing daily smokers to non-smokers/<daily smokers was 0.5 (i.e. a strong association between the exposure and missingness in the exposure); there was no bias when this odds ratio was 0.9 and only a small amount of bias when it was 0.75. The IPW estimates were slightly more biased than the MI estimates when two proxies were used and, as above, resulted in a loss of precision compared to the complete case analysis.

7.2.5.2.1 With an interaction

As described above (Section 7.2.5.1.1), the inclusion of the interaction resulted in large increases in bias (Table 7-18). The MI estimates when two linked smoking variables were used as auxiliaries were less biased (and more precisely estimated) than those obtained from the MI models with a single linked proxy (as well as those obtained in the complete case analysis). The IPW estimates with two proxies were less biased than the complete case estimates but more biased than the MI estimates if the model for the weights included interaction terms between the linked proxies and the outcome. If these interactions were excluded, the IPW estimates were very similar to those obtained when using only one proxy and, as above, were more biased than the complete case estimates. These results are shown in Table 7-18 and 7-19.

Table 7-16: Estimates of effect of daily smoking on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with two linked smoking variables: OR_{obs} comparing daily smokers to <daily/never = 0.75

	Factor 1: % missing data	Complete case analysis[1]		IPW using two linked proxies			MI using two linked proxies			
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision[2]	Estimate (empirical SE)	% bias	Gain in precision[2]	FMI[3]
Continuous outcome (true mean diff. =-0.8)	20%	-0.73 (0.040)	9%	-0.78 (0.044)	3%	-19%	-0.79 (0.038)	1%	13%	20%
	40%	-0.71 (0.049)	12%	-0.78 (0.060)	3%	-33%	-0.79 (0.043)	1%	30%	39%
	60%	-0.71 (0.065)	11%	-0.78 (0.087)	3%	-43%	-0.79 (0.055)	1%	42%	58%
	80%	-0.74 (0.103)	8%	-0.78 (0.145)	2%	-47%	-0.80 (0.079)	0%	71%	79%
Binary outcome (true log OR = 1.19)	20%	1.10 (0.086)	-11%	1.16 (0.084)	2%	-2%	1.18 (0.085)	-1%	1%	17%
	40%	1.08 (0.106)	-9%	1.16 (0.106)	3%	0%	1.18 (0.096)	-1%	24%	36%
	60%	1.09 (0.143)	-8%	1.16 (0.155)	2%	-15%	1.18 (0.122)	-1%	35%	58%
	80%	1.15 (0.225)	-4%	1.18 (0.248)	1%	-16%	1.19 (0.180)	0%	56%	80%

1. These are the same results as presented in Table 7-14 but are shown here to allow comparisons to be made with the corresponding MI results
2. Relative to complete case analysis
3. Fraction of missing information

Table 7-17: Estimates of effect of daily smoking on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with two linked smoking variables: OR_{obs} comparing daily smokers to <daily/never = 0.9 and 0.5

Factor 2: OR(obs) for smoking = 0.9	Factor 1: % missing data	Complete case analysis[1]		IPW with two linked proxies			MI using two linked proxies			
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision[2]	Estimate (empirical SE)	% bias	Gain in precision[2]	FMI[3]
Continuous outcome (true mean diff. =-0.8)	20%	-0.74 (0.039)	7%	-0.79 (0.043)	1%	-19%	-0.80 (0.037)	1%	10%	19%
	40%	-0.73 (0.048)	9%	-0.79 (0.058)	1%	-36%	-0.80 (0.042)	0%	28%	38%
	60%	-0.73 (0.065)	8%	-0.79 (0.081)	1%	-41%	-0.80 (0.053)	1%	51%	57%
	80%	-0.75 (0.098)	6%	-0.79 (0.134)	2%	-47%	-0.80 (0.073)	1%	77%	78%
Binary outcome (true log OR = 1.19)	20%	1.13 (0.084)	-5%	1.18 (0.084)	-1%	-2%	1.19 (0.084)	0%	1%	16%
	40%	1.12 (0.103)	-6%	1.18 (0.103)	-1%	-5%	1.19 (0.094)	0%	20%	35%
	60%	1.13 (0.138)	-5%	1.18 (0.146)	-1%	-11%	1.19 (0.117)	0%	39%	57%
	80%	1.17 (0.212)	-2%	1.19 (0.229)	0%	-16%	1.19 (0.171)	0%	54%	79%
OR(obs) for smoking = 0.5	% missing data	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision[2]	Estimate (empirical SE)	% bias	Gain in precision[2]	FMI[3]
Continuous outcome (true mean diff. =-0.8)	20%	-0.68 (0.042)	15%	-0.75 (0.048)	6%	-22%	-0.78 (0.040)	3%	11%	21%
	40%	-0.66 (0.056)	18%	-0.75 (0.068)	7%	-40%	-0.78 (0.048)	4%	30%	41%
	60%	-0.67 (0.079)	16%	-0.75 (0.102)	9%	-45%	-0.79 (0.060)	1%	58%	61%
	80%	-0.71 (0.121)	11%	-0.76 (0.177)	4%	-51%	-0.80 (0.087)	0%	95%	81%
Binary outcome (true log OR = 1.19)	20%	1.03 (0.090)	-14%	1.12 (0.090)	-6%	-2%	1.14 (0.088)	-4%	5%	18%
	40%	1.00 (0.117)	-16%	1.10 (0.121)	-7%	-6%	1.15 (0.102)	-4%	31%	38%
	60%	1.03 (0.167)	-13%	1.12 (0.183)	-6%	-16%	1.17 (0.136)	-1%	51%	60%
	80%	1.11 (0.269)	-7%	1.15 (0.300)	-3%	-20%	1.20 (0.201)	1%	69%	82%

1. These are the same results as presented in Table 7-15 but are shown here to allow comparisons to be made with the corresponding MI results
2. Relative to complete case analysis
3. Fraction of missing information

Table 7-18: Estimates of effect of daily smoking (vs <daily/never) on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with an interaction between the exposure and outcome with respect to missingness

(OR for interaction = 0.8)

	Factor 1: % missing data	Complete case analysis		IPW[1] using one linked proxy			MI using one linked proxy			
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision[2]	Estimate (empirical SE)	% bias	Gain in precision[2]	FMI[3]
Continuous outcome (true mean diff. =-0.8)	20%	-1.06 (0.037)	-31%	-1.07 (0.040)	-33%	-15%	-1.01 (0.034)	-26%	12%	20%
	40%	-1.21 (0.044)	-52%	-1.27 (0.053)	-59%	-29%	-1.10 (0.036)	-37%	46%	36%
	60%	-1.50 (0.053)	-87%	-1.54 (0.069)	-93%	-35%	-1.22 (0.037)	-53%	101%	50%
	80%	-1.54 (0.075)	-105%	-1.64 (0.133)	-105%	-67%	-1.14 (0.047)	-43%	149%	69%
Binary outcome (true log OR = 1.19)	20%	1.59 (0.081)	34%	1.59 (0.081)	33%	1%	1.54 (0.078)	30%	6%	20%
	40%	1.83 (0.099)	54%	1.86 (0.100)	56%	-3%	1.71 (0.093)	44%	13%	42%
	60%	2.30 (0.128)	93%	2.34 (0.130)	97%	-8%	2.00 (0.118)	68%	17%	63%
	80%	2.52 (0.159)	112%	2.38 (0.235)	100%	-52%	1.97 (0.148)	65%	15%	84%
		Complete case analysis		IPW[1] using two linked proxies			MI using two linked proxies			
Continuous outcome (true mean diff. =-0.8)	20%	As above		-0.99 (0.039)	-24%	-16%	-0.95 (0.035)	-19%	13%	17%
	40%			-1.15 (0.054)	-44%	-30%	-1.02 (0.036)	-28%	45%	33%
	60%			-1.38 (0.075)	-73%	-47%	-1.12 (0.042)	-40%	71%	50%
	80%			-1.32 (0.157)	-65%	-76%	-1.05 (0.050)	-31%	132%	68%
Binary outcome (true log OR = 1.19)	20%	As above		1.48 (0.082)	24%	0%	1.44 (0.080)	21%	4%	16%
	40%			1.68 (0.098)	41%	4%	1.57 (0.092)	28%	18%	35%
	60%			2.05 (0.135)	73%	-8%	1.78 (0.107)	50%	39%	57%
	80%			1.81 (0.228)	53%	-49%	1.73 (0.137)	46%	39%	79%

1. In which interaction term(s) (between the linked proxy/ies and the continuous outcome were included in the model for the weights
2. Relative to complete case analysis
3. Fraction of missing information

Table 7-19: IPW estimates of effect of daily smoking on attainment score (continuous outcome) and not obtaining five or more A*- C grades (binary outcome) with an interaction between the exposure and outcome with respect to missingness

	Factor 1: % missing data	IPW[1] using one linked proxy		
		Estimate (empirical SE)	% bias	Gain in precision[2]
Continuous outcome (true mean diff. =-0.8)	20%	-1.09 (0.039)	-37%	-13%
	40%	-1.32 (0.052)	-65%	-26%
	60%	-1.60 (0.065)	-100%	-27%
	80%	-1.79 (0.082)	-124%	-12%
Binary outcome (true log OR = 1.19)	20%	1.63 (0.082)	37%	-2%
	40%	1.95 (0.101)	64%	-5%
	60%	2.47 (0.129)	108%	-6%
	80%	2.75 (0.175)	131%	-13%
		IPW[1] using two linked proxies		
Continuous outcome (true mean diff. =-0.8)	20%	-1.09 (0.040)	-36%	-17%
	40%	-1.31 (0.053)	-64%	-28%
	60%	-1.60 (0.067)	-100%	-33%
	80%	-1.79 (0.080)	-123%	-10%
Binary outcome (true log OR = 1.19)	20%	1.62 (0.082)	36%	-1%
	40%	1.93 (0.100)	55%	-2%
	60%	2.46 (0.135)	106%	-8%
	80%	2.72 (0.172)	129%	-11%

1. In which interaction term(s) (between the linked proxy/ies and the continuous outcome were excluded from the model for the weights
2. Relative to complete case analysis

7.3 Discussion

In Section 7.1.3 I discussed the results of the exemplar and suggested that I would have expected the exposure-outcome estimates from MI to be less biased than those obtained from the complete case analysis. However, in many cases the actual estimates were quite similar – perhaps because the bias in the different analyses was largely similar.

In this exemplar I had three different (GP-recorded) proxies for smoking: ever-smoked, current smoking, and smoking aged 16-19. Thus, the simulations with two linked proxies possibly match the exemplar most closely. Using the available data, it was not possible to rule out there being an interaction between smoking (using the

GP measures of smoking) and educational attainment (exposure and outcome) with respect to missingness, particularly for frequency of smoking. Given all this, and based on the results of the simulations, I would have expected greater differences between the observed estimates for the association between frequency of smoking and the attainment score. It seems likely that these estimates are all biased to some degree. For the binary outcome (not obtaining five or more A*- C grades), the simulations suggest that the odds ratio for frequency of smoking in the MI analysis is less biased than that obtained from the complete cases analysis, but probably still an over-estimate.

Clearly, the simulations do not exactly replicate the real situation. Key differences include the fact that I also had missing data in the linked proxies, and in both the outcome and covariates (in addition to the exposures of interest). Having some missing data in the linked proxies would probably result in smaller improvements in terms of bias reduction and gains in precision compared to a situation in which linked data were available for the whole sample.

In conclusion, the simulations suggest that, when you have a binary exposure that is MNAR – and missingness in this exposure is also dependent on the outcome of interest – incorporating linked proxies in MI can reduce bias and increase precision, particularly when there is more than one proxy (or potentially a proxy with high sensitivity and specificity, although I did not simulate this). I would not recommend using IPW since the estimates from IPW were less precise than the complete case estimates and did not result in greater reductions in bias than MI. In addition, in some situations (where there was an interaction between the exposure and outcome with respect to missingness but an interaction was not included in the model for the weights) IPW produced estimates that were more biased (as well as being less precise) than the complete case estimates.

If missingness were dependent on the exposure but not the outcome then a complete case analysis would produce unbiased estimates of the exposure-outcome association. It could be argued that the linked variables would not add anything in

this scenario – although if they were strongly associated with the exposure then you would get increases in efficiency. However, having one or more linked proxies for the missing exposure allows the missingness mechanism to be more fully investigated. For example, in the above exemplar it allowed me to examine whether there was likely to be an interaction between smoking and educational attainment with respect to missingness. Since the missingness mechanism determines which analysis is likely to be the most appropriate, obtaining proxies for the missing variables is advantageous regardless of whether or not they are used in the final analysis.

Chapter 8 Misclassification in a binary exposure

In this chapter I use a gold standard measure to examine the impact of misclassification (on exposure-outcome estimates) in a binary exposure. I revisit the exemplar study described in Chapter 7 in which the aim was to estimate the impact of teenage smoking on educational attainment at age 16 years. As described in Chapter 7, I have two outcomes of interest: a continuous outcome (attainment score) and a binary outcome (whether an individual did not or did obtain five or more A* to C grades at GCSE: 1=did not obtain 5+ A*- C grades, 0=did). In Chapter 7 I had three measures of teenage smoking (at age 15 years): ever-smoked, frequency of current smoking, and – on a subset of individuals – cotinine, dichotomised to classify individuals as daily smokers or not. In this chapter I focus on frequency of current smoking, coded (as in Chapter 7) as a binary variable measuring whether or not an individual was a daily smoker. Thus, the two analysis models of interest are a multiple linear regression of the attainment score on teenage smoking (and other covariates) and a multiple logistic regression for the binary outcome.

In an ideal world we would have a perfect measure of teenage smoking. However, in this case, the main exposure measure is self-reported frequency of smoking; this is almost certainly subject to misclassification. I also have current smoking as recorded in the GP data. This is also likely to be subject to misclassification, particularly as many teenagers may not visit their GP regularly at this age, so the GP records will potentially only "pick up" a relatively small proportion of smokers. As such, we would expect the estimates of the association between smoking and attainment – whether using self-reported or GP-recorded smoking – to be biased.

Although cotinine is not a perfect measure of daily smoking, it is considered a gold standard (Jarvis et al. 1987). I use this to examine misclassification in both the self-reported and the GP-recorded smoking data and to investigate different methods for correcting for misclassification in the presence of a gold standard measure. After analysing the ALSPAC data I also carry out a small set of simulations based on this example.

8.1 Analysis

Initially I carried out a descriptive analysis using the gold standard measure (cotinine) to examine misclassification in both self-reported smoking and GP-recorded smoking. I then corrected for misclassification in several different ways: using probabilistic bias analysis, multiple imputation, and a Bayesian analysis. In all cases, I initially disregarded covariates and obtained corrected estimates of the crude association between teenage smoking and educational attainment. These analyses were restricted to those with non-missing outcomes and non-missing self-reported and/or GP-recorded smoking status.

The analyses involving GP-recorded smoking were carried out twice, once including all individuals for whom GP-recorded smoking data were available and once restricting to individuals who attended the clinic during which cotinine was measured. The latter was carried out because individuals who did not have cotinine measured were a combination of (i) individuals who did not attend the clinic in which it was measured and (ii) individuals who attended the clinic but did not have/want a sample taken. The factors influencing missingness in cotinine among those who attended the clinic may be different to the factors influencing attendance at the clinic and these factors are also likely to influence the association between GP-recorded smoking and cotinine-measured smoking.

I then expanded the analyses to obtain corrected values of the fully adjusted estimates of the association between smoking and attainment (both the continuous outcome and the binary outcome). For the fully adjusted analyses (adjusting for the

same set of covariates described in Chapter 7: child's sex and KS2 attainment score; maternal age, parity, marital status, depression score during pregnancy, ethnicity and smoking; mother's and father's educational level; paternal smoking (ever smoked); family occupational social class; housing tenure; use of car; phone in home; double glazing; number of rooms in the house and financial difficulties score), all analyses were additionally restricted to individuals with fully observed covariates.

8.1.1 Probabilistic bias analysis (PBA)

This method is outlined in Chapter 2 and, as described previously (Section 2.2.1), involves three main steps:

Step 1: Modelling the bias parameters

For both self-reported and GP-recorded smoking, the observed cell counts from a cross-tabulation with cotinine, separated by outcome status (attainment: 5 or more A*- C grades / not) (Table 8-3) were used to specify probability distributions for the positive and negative predictive values (PPV and NPV) among those who did and those who did not obtain five or more A* to C grades at GCSE. This was to take account of differential misclassification of the exposure (smoking) by the outcome (attainment). These were specified as beta distributions with means given by the observed PPVs / NPVs. So, the four parameters for self-reported smoking were given the following probability distributions:

PPV_{5+} ~ beta(#true positives$_{(5+)}$, #false positives$_{(5+)}$)

NPV_{5+} ~ beta(#true negatives$_{(5+)}$, #false negatives$_{(5+)}$)

$PPV_{not\ 5+}$ ~ beta(#true positives$_{(not\ 5+)}$, #false positives$_{(not\ 5+)}$)

$NPV_{not\ 5+}$ ~ beta(#true negatives$_{(not\ 5+)}$, #false negatives$_{(not\ 5+)}$)

where PPV_{5+} indicates the positive predictive value among those who received five or more A*- C grades, and so on. For GP-recorded smoking, the corresponding values from the lower half of Table 8-3 were used. Estimates of the PPV and NPV were then sampled from these distributions (separately by outcome status).

Step 2: Correction for misclassification

Following this, these sampled values were applied to the individual data points to simulate a person's true (corrected) smoking status, given their (binary) outcome and observed (self-reported or GP-recorded) smoking status. These were simulated from a Bernoulli distribution with the sampled estimate of the NPV and PPV, as appropriate. For each simulation, the log odds ratio (for the binary outcome, not obtaining five or more A*- C grades) and regression coefficient (for the continuous outcome, attainment score) for daily smoking was calculated using the corrected smoking variable. This gave one log odds ratio (for the binary outcome) or regression coefficient (for the continuous outcome) based on the sampled values of the PPV and NPV.

Step 3: Estimate the parameter of interest

The above process – the final part of step 1 (sampling the NPV and PPV) and step 2 – was repeated 10,000 times and the resulting estimates (log odds ratio for the binary outcome and regression coefficient for the continuous outcome) were saved. As described in Chapter 2, to take account of the sampling error introduced in the estimation of the uncorrected estimates, I multiplied the observed standard error in the original (naïve) analysis by a randomly chosen value from the standard normal distribution then subtracted this product from the observed log odds ratio / regression coefficient.

8.1.2 Multiple imputation

In this analysis, true smoking status (daily vs. less than daily or never) was classified as observed (and equal to smoking status according to their cotinine level) for all individuals among whom it was measured. Among individuals for whom cotinine was not measured, true smoking status was classified as missing. True smoking status was then imputed using a logistic model. To take account of differential misclassification, I carried out the imputations separately among those who did and did not obtain five or more A* to C grades (i.e. stratified by the binary outcome). The imputation models included self-reported smoking and/or GP-recorded smoking, the continuous

outcome (for the analysis using the binary outcome, I carried out two separate sets of imputations: one in which the continuous outcome was included in the imputation model and one in which it was excluded), and, in the adjusted analysis, measured covariates. Note that covariates could also have been included in the MI model for the unadjusted analysis. However, to make the crude analyses more comparable (i.e. with true smoking dependent only on measured smoking and attainment in each analysis) I did not do this. As described above, in the fully adjusted analysis, each subset was additionally restricted to those with fully observed covariates because I wanted to focus the analysis on only correcting for misclassification. I carried out this analysis in five subsets of individuals:

1. All those with observed values for self-reported smoking and outcomes (binary and continuous).
2. All those with observed values for GP-recorded current smoking and outcomes.
3. As above (in 2.), but additionally restricting to those who attended the clinic at which cotinine was measured.
4. All those with either self-reported or GP-recorded current smoking and complete outcome data.
5. As above (in 4.), but additionally restricting to those who attended the clinic at which cotinine was measured.

In each case, 100 datasets were imputed with a burn-in of 10 iterations. Stata's *mi impute* command was used to carry out the imputations.

If the gold standard (assumed in this case to be the true smoking status) had been obtained on a random sample of all individuals and misclassification of both self-reported and GP-recorded smoking was the same regardless of whether these data were missing, each of the above five analyses would be expected to result in approximately the same corrected estimates. Therefore, I carried out these separate analyses to highlight the potential (additional) impact of missing data on the estimates of the association between smoking and attainment.

Further, to highlight the potential impact of missing data (the fact that true smoking was not simply available for a random subset of individuals), I also present results from Chapter 7 showing the estimate of the effect of cotinine on the attainment score and the binary outcome obtained using MI on the complete sample. In this analysis the outcomes and covariates were also imputed using information on all other exposures, including smoking measures derived from the GP records (ever smoked as well as future smoking). This analysis would simultaneously correct for missing data and misclassification.

8.1.3 Bayesian analysis

As described in Section 2.2.3, this involves specifying three sub-models:

The outcome model

The outcome (attainment) was modelled as conditional on true smoking status (unknown or missing for those without cotinine measured) and, in the adjusted analysis, also on the covariates specified above. This was a logistic model for the binary outcome (5+ A*-C grades: no/yes) and a normal linear regression model for the continuous attainment score. Non-informative normal priors [$N(0, 100^2)$] were placed on all the parameters in these models.

The measurement (misclassification) model

In this part of the model, self-reported (or GP-recorded) smoking was specified as dependent on an individual's true smoking status and their outcome status (did not / did obtain five or more A*- C grades):

Self-reported smoking ~ Bernoulli(p_{ij}), for i=0,1 and j=0,1

where i represents the individual's true smoking status (as measured by cotinine; 0=non-smoker, 1=smoker) and j represents their outcome status.

The p_{ij}s represent the sensitivity and false positive rate (1-specificity) of self-reported smoking among those who did and did not obtain five or more A*- C grades. In the Bayesian analysis these are treated as random variables to be estimated and

information from this part of the model feeds into the analysis model and vice versa. Non-informative priors [Uniform(0,1)] were also placed on these parameters. This analysis was carried out on the same five subsets of individuals as specified in Section 8.1.2 above.

Note that in the analysis including both self-reported and GP-recorded smoking I specified:
Self-reported smoking ~ Bernoulli(p_{ij}), for i=0,1 and j=0,1
GP-recorded smoking ~ Bernoulli(q_{ij}), for i=0,1 and j=0,1
with i and j as specified above.

The exposure model

Initially I simply modelled true smoking as following a Bernoulli distribution with probability φ. This was given a non-informative prior: Uniform(0,1). I later extended this so that true smoking was modelled (using logistic regression) as a function of the child's sex and KS2 attainment score as well as maternal and paternal smoking. These were the main predictors of cotinine in the ALSPAC data.

OpenBUGS (http://www.openbugs.net) was used to carry out the Bayesian analysis. For each sub-analysis (1-5 listed above), two Markov chains were run. For the crude analysis, I ran 25,000 iterations for each chain, 5,000 of which were discarded as burn-in. For the adjusted analysis, I had a burn-in of between 50,000 and 150,000 iterations followed by 20,000 iterations for each chain. Different burn-ins were used because the models with only self-reported smoking converged more quickly than those that included GP-recorded smoking. Note that the second Bayesian model (in which true smoking was modelled as a function of sex, KS2 attainment, and maternal and paternal smoking) did not run (OpenBUGS crashed repeatedly). Thus, the results presented in this chapter used the very simple exposure model.

8.2 Results

The numbers with observed data for the different exposures, as well as numbers of these with outcome data and fully observed covariates, are given in Table 8-1. A total of 3,441 individuals had cotinine (the gold standard) measured at age 15 years. Altogether 5,328 individuals had self-reported data on frequency of smoking at age 15 and 9,521 had GP-recorded current smoking at age 15. Among these, 3,351 (63%) and 2,573 (27%) – respectively – also had cotinine measured.

As a reminder, the capped attainment score (the continuous outcome in this analysis) ranged from 0 to 540, with overall mean (SD) of 315 (96). Among individuals who attended the clinic at which cotinine was measured, the mean was higher (350, SD=75 points). Overall, 59% of individuals obtained five or more A*-C grades (binary outcome); again, this was higher among those who attended the clinic when cotinine was measured (76%).

Table 8-1: Numbers with available data for exposures (smoking variables), outcomes (attainment), and covariates and number (%) of these with gold standard (cotinine) data

Smoking variable (exposure)	Exposure measured? = Yes	Outcome measured?	N	Among these, number (%) with cotinine measured	Exposure, outcome and covariates measured? = Yes	N	Among these, number (%) with cotinine measured
Cotinine		Yes	2,989	All		1,682	All
		No	452			--[2]	--[2]
		Total[1]	3,441			--[2]	--[2]
Self-report		Yes	4,669	2,911 (62%)		2,596	1,650 (64%)
		No	659	--[2]		--[2]	--[2]
		Total[1]	5,328	3,351 (63%)		--[2]	--[2]
GP-recorded (of whom attended clinic)		Yes	8,402 (3,731)	2,311 (28%) (2,331, 62%)		3,846 (2,096)	1,324 (34%) (1,324, 63%)
		No	1,119	--[2]		--[2]	--[2]
		Total[1]	9,521 (4,115)	2,573 (27%) (2,573, 63%)		--[2]	--[2]
Either self-report or GP (of whom attended clinic)		Yes	9,432 (4,769)	2,967 (31%) (2,967, 62%)		4,380 (2,630)	1,671 (38%) (1,671, 64%)
		No	1,404	--[2]		--[2]	--[2]
		Total[1]	10,836 (5,430)	3,412 (31%) (3,412, 63%)		--[2]	--[2]

1. Total with this exposure variable measured
2. These numbers are not given

8.2.1 Misclassification

Table 8-2 shows the relationship between both self-reported and GP-recorded smoking and cotinine. Both had high specificity (99%). However, self-reported smoking correctly identified just under two thirds of individuals as daily smokers but GP-recorded smoking only correctly identified 22%.

Table 8-2: Comparison of self-reported and GP-recorded smoking with cotinine

		Cotinine measured at age 15 (gold standard)	
		≤9.5 ng/ml	>9.5 ng/ml
Self-reported smoking	Never/<daily Daily	3,034 (99%) 30	100 187 (65%)
GP-record smoking[1]	No/uncertain[2] Yes	2,317 (99%) 22	182 52 (22%)

1. Current smoking at 15 years
2. Uncertain smoking status – smoking cessation codes but not positive record of being a current smoker OR smoking recorded with missing value for amount smoked. Only 12 individuals fell into this category and others (Atkinson et al. 2017) have classified the latter group as non-smokers; for these reasons, I combined uncertain with non-smokers

Table 8-3 shows the relationship between the smoking variables separately by outcome status (did not/did obtain five or more A* to C grades at GCSE) – to examine whether there was evidence for differential misclassification in smoking. The specificity of both self-reported and GP-recorded smoking was slightly higher among those who obtained five or more A*- C grades: 99.5%; 95% CI (99.1, 99.7%) for self-reported smoking and 99.3% (98.8, 99.7%) for GP-recorded smoking compared to those who did not obtain five or more A*- C grades: 96.8% (94.9, 98.2) for self-reported smoking and 97.5% (95.5, 98.8) for GP-recorded smoking. However, the sensitivity was lower among those who obtained five or more A*- C grades compared to those who did not for both self-reported and GP-recorded smoking and, as previously, higher for self-reported compared to GP-recorded smoking: 55.9% (46.5, 65.1) for those who obtained five or more A*- C grades and 69.7% (61.5, 77.0) for those who not (for self-reported smoking) compared to 15.5% (8.9, 24.2) and 27.9% (20.1, 36.7) for GP-recorded smoking.

Table 8-3: Comparison of self-reported and GP-recorded smoking with cotinine by outcome status

		Obtained five or more A*-C grades?			
		Yes		No	
		Cotinine			
		≤9.5 ng/ml	>9.5 ng/ml	≤9.5 ng/ml	>9.5 ng/ml
Self-reported smoking	Never/<daily Daily	2,134 (99.5%) 11	52 66 (55.9%)	487 (96.8%) 16	44 101 (69.7%)
GP-recorded smoking[1]	No/uncertain[2] Yes	1680 (99.3%) 11	82 15 (15.5%)	391 (97.5%) 10	88 34 (27.9%)

1. Current smoking at 15 years
2. Uncertain smoking status – smoking cessation codes but not positive record of being a current smoker OR smoking recorded with missing value for amount smoked

8.2.2 Taking account of misclassification: comparison of methods

8.2.2.1 Crude analysis

The estimates of the association between teenage smoking and educational attainment for both the binary and continuous outcome are given in Table 8-4. This includes the estimates from the naïve analyses (i.e. not correcting for misclassification) as well as the corrected estimates.

For the binary outcome (did not obtain 5+ A* - C grades vs did), the naïve analyses using self-reported or GP-recorded smoking resulted in over-estimates of the association between smoking and educational attainment (i.e. compared to the analysis using the gold standard measure). All three methods for taking account of misclassification resulted in quite similar estimates of the log odds ratio, particularly when restricting to those who attended the clinic at which cotinine was measured and, in the MI models, when the continuous outcome was excluded from the imputation models when estimating the effect of smoking on the binary outcome. Not restricting to this subset of individuals resulted in slightly larger estimates of the log odds ratio. When the continuous outcome was included in the MI models to estimate the effect of smoking on the binary outcome, some of the results were slightly different from those in which it was excluded; this was particularly the case when only GP-recorded smoking was being used to impute missing (true) smoking status.

For the continuous outcome, the attainment score, there were greater differences in the corrected estimates. Probabilistic bias analysis resulted in values closer to the null compared to the estimate based on the gold standard smoking measure, whereas the Bayesian analysis resulted in more extreme estimates. The estimates from MI – when restricting to individuals who attended the clinic at which cotinine was measured – were quite close to the estimate based on the gold standard. Not restricting to the subset of individuals who attended the clinic at which cotinine was measured resulted in more extreme estimates of the mean difference in attainment score, particularly for the Bayesian analysis.

Table 8-4: Effect of smoking at 15 years on educational attainment at 16: comparison of methods to take account of misclassification: crude estimates
(log odds ratio for binary outcome and regression coefficient for continuous outcome)

Analysis method (all are crude effect estimates)	N	Did not obtain 5+ A*- C grades	KS4 attainment score
Naive analyses			
Cotinine (gold standard) at 15 years (>9.5 vs ≤9.5 ng/ml)	2,989	1.68 (1.43, 1.94)	-83 (-91, -74)
Self-reported smoking at 15 years (daily vs < daily/never)	4,699	1.83 (1.59, 2.06)	-85 (-92, -77)
GP-recorded smoking at 15 years (yes vs no)	8,402	1.72 (1.51, 1.94)	-94 (-103, -86)
GP-recorded smoking at 15 years (yes vs no); restricted to those who attended clinic when cotinine was measured	3,731	1.78 (1.40, 2.16)	-93 (-106, -80)
Probabilistic bias analysis, corrected values of:			
Self-reported smoking	4,699	1.66 (1.41, 1.91)	-74 (-82, -65)
GP-recorded smoking	8,402	1.68 (1.44, 1.92)	-73 (-83, -63)
GP-recorded smoking, restricted	3,731	1.66 (1.27, 2.06)	-74 (-88, -59)
Multiple imputation			
Continuous score included in imputations			
Using self-report only	4,699	1.64 (1.41, 1.88)	-79 (-87, -70)
Using GP records only	8,402	1.88 (1.62, 2.14)	-117 (-131, -104)
Using GP records only, restricted	3,731	1.67 (1.39, 1.95)	-85 (-96, -75)
Using self-report & GP records	9,432	1.74 (1.50, 1.98)	-97 (-110, -85)
Using self-report & GP records, restricted	4,761	1.66 (1.43, 1.89)	-80 (-88, -72)
Continuous score excluded from imputations			
Using self-report only	4,699	1.66 (1.43, 1.89)	N/A
Using GP records only	8,402	1.67 (1.38, 1.97)	N/A
Using GP records only, restricted	3,731	1.66 (1.38, 1.94)	N/A
Using self-report & GP records	9,432	1.70 (1.47, 1.92)	N/A
Using self-report & GP records, restricted	4,761	1.67 (1.44, 1.89)	N/A
Bayesian analysis			
Using self-report only	4,699	1.66 (1.42, 1.89)	-93 (-101, -85)
Using GP records only	8,402	1.68 (1.43, 1.94)	-179 (-184, -174)
Using GP records only, restricted	3,731	1.66 (1.38, 1.95)	-105 (-114, -96)
Using self-report & GP records	9,432	1.68 (1.47, 1.90)	-168 (-173, -163)
Using self-report & GP records, restricted	4,761	1.65 (1.43, 1.87)	-94 (-101, -85)

8.2.2.2 Adjusted analysis

For the fully adjusted analysis, there were 2,596 individuals with complete covariates, outcome and self-reported smoking data; 1,650 (64%) of these had cotinine measured. Similarly, there were 3,846 individuals with complete covariates, outcome and GP-recorded smoking data, of whom 1,324 (34%) had cotinine measured. Finally, 4,380 individuals had complete covariates, outcome data and either self-reported or GP-recorded data on smoking; among these 1,671 (38%) had cotinine measured.

For the binary outcome (did not obtain 5+ A* - C grades vs did), some of the models that included GP-recorded smoking failed to reach convergence after 180,000 iterations: one chain appeared to have converged to a stationary distribution but the other chain had not. I have presented the results from the former chain only but cannot be sure that the these have been sampled from the true posterior distribution. For the continuous outcome (attainment score), all the models that included GP-recorded smoking updated very slowly and OpenBUGS crashed repeatedly; I was only able to get two of them to run. I think these issues might have arisen because of the small numbers with a positive record (i.e. classified as a smoker) for GP-recorded smoking.

Table 8-5 shows the fully adjusted estimates of the association between teenage smoking and the two outcomes using the different methods to take account of misclassification. There were greater differences between the estimates – for both outcomes – in the fully adjusted analysis. The corrected log odds ratio for not obtaining five or more A* to C grades varied from 1.47 to 2.12 and, for the continuous outcome, the estimates ranged from -19 to -69. Using PBA still appeared to result in estimates of the regression coefficient for smoking on the attainment score that were closer to the null than estimates from the other methods. The Bayesian estimates that I was able to obtain were further from the null than the MI estimates but the differences were smaller than those seen in the crude analysis.

Table 8-5: Effect of smoking at 15 years on educational attainment at 16 years: comparison of methods to take account of misclassification: fully adjusted estimates (log odds ratio for binary outcome and regression coefficient for continuous outcome)

Analysis method (all are fully adjusted effect estimates)	N	Did not obtain 5+ A*- C grades	KS4 attainment score
Naive analyses			
Cotinine (gold standard) at 15 years (>9.5 vs ≤9.5 ng/ml)	1,682	1.59 (1.05, 2.12)	-42 (-49, -35)
Self-reported smoking at 15 years (daily vs < daily/never)	2,068	2.12 (1.66, 2.58)	-51 (-57, -44)
GP-recorded smoking at 15 years (yes vs no)	3,846	1.87 (1.38, 2.37)	-60 (-68, -51)
GP-recorded smoking at 15 years, restricted to those who attended clinic at which cotinine was measured	2,096	1.93 (1.10, 2.76)	-51 (-66, -51)
Probabilistic bias analysis, corrected values of:			
Self-reported smoking	2,596	1.79 (1.28, 2.30)	-39 (-43, -36)
GP-recorded smoking	3,846	1.67 (1.12, 2.21)	-34 (-44, -24)
GP-recorded smoking, restricted	2,096	1.61 (0.75, 2.47)	-36 (-49, -23)
Multiple imputation			
Continuous score included in imputations			
Using self-report only	2,596	1.73 (1.23, 2.24)	-42 (-50, -35)
Using GP records only	3,846	1.76 (1.21, 2.29)	-69 (-79, -58)
Using GP records only, restricted	2,096	1.61 (1.02, 2.20)	-47 (-56, -38)
Using self-report and GP records	4,380	2.12 (1.66, 2.58)	-55 (-66, -44)
Using self-report and GP records, restricted	2,630	1.93 (1.43, 2.43)	-43 (-50, -36)
Continuous score excluded from imputations			
Using self-report only	2,596	1.91 (1.39, 2.43)	N/A
Using GP records only	3,846	1.57 (0.99, 2.15)	N/A
Using GP records only, restricted	2,096	1.59 (0.99, 2.19)	N/A
Using self-report and GP records	4,380	2.06 (1.60, 2.52)	N/A
Using self-report and GP records, restricted	2,630	1.91 (1.41, 2.42)	N/A
Bayesian analysis			
Using self-report only	2,596	1.82 (1.34, 2.31)	-50 (-57, -44)
Using GP records only	3,846	1.47 (0.92, 2.00)	[2]
Using GP records only, restricted	2,096	1.63 (1.04, 2.24)[1]	-55 (-62, -48)[3]
Using self-report and GP records	4,380	1.70 (1.26, 2.16)[1]	[2]
Using self-report and GP records, restricted	2,630	1.78 (1.30, 2.24)[1]	[2]

1. One chain had not converged to a stationary distribution after 180,000 iterations but the other chain appeared to have reached convergence. These are the results from the second chain only.
2. These models would not update in OpenBUGS: they crashed repeatedly
3. Based on only 1000 iterations after 1000 burn-in – after this, OpenBUGS crashed repeatedly. The chains did appear to have converged to a stationary distribution but the MC error was obviously higher than it would otherwise have been

Finally, Table 8-6 shows the estimated association between cotinine and both the binary and continuous outcome obtained using multiple imputation on the whole sample (all ALSPAC enrolled singletons and twins, alive at one, not withdrawn from the study and with a valid NHS number, as described in Chapter 3). In this analysis,

and as described in Chapter 7, the MI model included both outcomes, the exposure (cotinine), all self-reported and two GP-recorded smoking measures, as well as the covariates listed in Section 8.1; the imputations were carried out separately by sex (but not separately by the binary outcome).

Table 8-6: Estimates (log odds ratio and regression coefficient) of the effect of smoking at 15 years on educational attainment at 16 years: MI on complete sample

Exposure	N	Did not obtain 5+ A*-C grades	KS4 attainment score
Cotinine ≥9.5 vs <9.5 ng/ml	14,566	1.41 (1.10, 1.73) [OR = 4.11 (3.01, 5.62)]	-52 (-60, -44)

8.3 Simulations

I carried out simulations in order to assess the extent to which (i) the results above might be being affected by missing data and (ii) the different correction methods are likely to reduce bias in different scenarios. I simulated ten datasets of 100,000 observations each containing the following variables with distributions and associations similar to those seen in the ALSPAC data:

- The **continuous attainment score** (KS4) was simulated as a normal random variable with mean 315 and standard deviation 75 points. If an individual's score was simulated as being negative then it was set to zero.

- The **probability of NOT obtaining five or more A*- C grades** (ks4not5) was generated from a logistic model dependent on the continuous attainment score:

  $\text{logit}(\Pr(\text{ks4not5})_i) = 24 - (0.082 \times \text{KS4}_i)$ where i represents an individual

- **True smoking status** (smoke) was simulated as two Bernoulli random variables with probability 21% if an individual did not obtain five or more A*- C grades and 5% if they did.

- **Self-reported smoking** (SRsmoke) was simulated using a logistic model, again based on relationships and distributions seen in the ALSPAC data:

  $\text{logit}(\Pr(\text{SRsmoke})_i) = -3.2 - (0.005 \times \text{KS4}_i) + (5.3 \times \text{smoke}_i) + (1.15 \times \text{ks4not5}_i) - (1.25 \times \text{smoke}_i \times \text{ks4not5}_i)$

I simulated two types of dataset (two sets of five datasets each) with different **mechanisms for inclusion in the validation subsample**:

- A random sample of 60% of all individuals – i.e. 40% were simulated to have missing data on true smoking (i.e. MCAR)

- True smoking was generated as MNAR (again, with 40% missing), with the missingness mechanism dependent on both outcomes (continuous and binary) using the following logistic model:

  $\text{logit}(\Pr(\text{smoke\_miss}))_i = 1.1 + (0.4 \times \text{smoke}_i) - (0.005 \times \text{KS4}_i) + (0.1 \times \text{ks4not5}_i) - (0.06 \times \text{smoke}_i \times \text{ks4not5}_i)$

Note that in the former situation – where true smoking is MCAR – the estimate of the effect of true smoking on attainment obtained using the subsample will be unbiased whereas, in the latter, I generated it such that the estimates for both the continuous and binary outcome would be biased (i.e. as shown in the logistic model above, I made missingness in true smoking dependent on both outcomes, true smoking status, and the interaction between true smoking status and the binary outcome).

8.3.1 Analysis

I carried out the same analyses as described above, but in the Bayesian analyses I only ran 6,000 iterations for each chain (with 1,000 discarded as burn-in) because the models converged very quickly. In the MI models, as above, true smoking was imputed using a logistic model; the models included self-reported smoking and the continuous outcome, and were stratified by the binary outcome.

8.3.2 Results

8.3.2.1 True smoking MCAR

The results for the first dataset in which true smoking was MCAR are shown in Table 8-7 and for the remaining four datasets in Appendix B, Table 14. As in the analysis of the ALSPAC data, estimates of the effect of smoking on educational attainment using the misclassified measure of smoking were biased away from the null compared to the estimates obtained using the gold standard (true) smoking status. All correction methods gave quite similar estimates of the log odds ratio. In some cases, these corrected estimates were very close to the value seen in the full dataset (datasets 1 to 3) whereas in others they were either slight under-estimates (dataset 4) or over-estimates (dataset 5).

The correction methods also gave reasonably similar estimates for the regression coefficient for the continuous attainment score, although the estimate corrected using PBA was slightly biased away from the null. The Bayesian estimates were, on average, slightly higher than the MI estimates, but this did not hold in all cases.

Table 8-7: Effect of smoking on educational attainment: comparison of methods to take account of misclassification: estimates (log odds ratio and regression coefficient for smoking) in simulated dataset of 100,000 observations with true smoking MCAR

Analysis method (all are crude effect estimates)	N	Did not obtain 5+ more A*- C grades	KS4 attainment score
Naive analyses			
True smoking, complete	100,000	1.662 (1.618, 1.707)	-44.7 (-46.2, -43.3)
True smoking, observed	59,726	1.675 (1.617, 1.733)	-45.0 (-46.9, -43.1)
Self-reported smoking	100,000	1.775 (1.724, 1.826)	-52.1 (-53.7, -50.5)
Methods used to correct for misclassification			
Probabilistic bias analysis	100,000	1.670 (1.614, 1.725)	-46.3 (-48.0, -44.6)
Multiple imputation (continuous included[1])	100,000	1.668 (1.615, 1.722)	-45.0 (-46.8, -43.3)
Multiple imputation (continuous excluded[1])	100,000	1.670 (1.618, 1.721)	N/A
Bayesian analysis	100,000	1.672 (1.621, 1.720)	-45.4 (-47.1, -43.7)

1. In imputation models when effect of smoking on binary outcome was being estimated

8.3.2.2 True smoking MNAR

The results obtained in the first simulated dataset when true smoking was MNAR are given in Table 8-8. Making true smoking MNAR resulted in a small amount of bias in both the regression coefficient for the continuous outcome and the log odds ratio for the binary outcome; in both cases the bias was towards the null. As when true smoking was MCAR, using the misclassified measure of smoking resulted in bias away from the null for both outcomes. The bias due to misclassification was greater in magnitude than the bias due to missing data in true smoking. The same pattern was observed in the other datasets (Appendix B, Table 15).

The correction methods again gave very similar results for the binary outcome (Table 8-8 and Appendix B, Table 15). In all cases, these estimates were between the value seen in the full dataset and the value estimated from complete cases (recall that this was slight biased towards the null due to missing data (MNAR) in true smoking). The regression coefficient for the continuous outcome (attainment score) obtained using PBA was again slightly biased away from the null. The MI estimates were, as with the binary outcome, all between the value seen in the full dataset and the estimate obtained from complete cases (true smoking observed). Finally, there also appeared to be some bias (away from the null) in the Bayesian estimates, although these were all closer to the value seen in the full dataset than the estimates obtained using PBA.

Table 8-8: Effect of smoking on educational attainment: comparison of methods to take account of misclassification: estimates (log odds ratio and regression coefficient for smoking) in simulated dataset of 100,000 observations with true smoking MNAR

Analysis method (all are crude effect estimates)	N	Did not obtain 5+ more A*-C grades	KS4 attainment score
Naive analyses			
True smoking, complete	100,000	1.623 (1.579, 1.667)	-44.2 (-45.7, -42.8)
True smoking, observed	59,053	1.575 (1.514, 1.635)	-42.6 (-44.6, -40.4)
Self-reported smoking	100,000	1.722 (1.672, 1.772)	-51.7 (-53.3, -50.1)
Methods used to correct for misclassification			
Probabilistic bias analysis	100,000	1.607 (1.552, 1.662)	-45.6 (-47.4, -43.9)
Multiple imputation (continuous included[1])	100,000	1.607 (1.554, 1.661)	-44.1 (-45.9, -42.3)
Multiple imputation (continuous excluded[1])	100,000	1.608 (1.554, 1.662)	N/A
Bayesian analysis	100,000	1.606 (1.553, 1.660)	-45.5 (-47.2, -43.7)

1. In imputation models when effect of smoking on binary outcome was being estimated

8.4 Discussion

In the ALSPAC data the crude results for the binary outcome were all reasonably similar, particularly when restricted to those who attended the clinic at which cotinine was measured. However, there were greater differences between the corrected crude estimates for the continuous outcome and the fully adjusted estimates for both the continuous and binary outcome. My simulations clearly did not match the ALSPAC data in all ways. The differences and their implications in terms of the exemplar are discussed below.

Misclassified exposures incomplete

In the simulations above I did not induce missingness in the misclassified exposure. In contrast, in the ALSPAC data, as well as cotinine (the gold standard) being missing on a large proportion of individuals, there was also missing data in both the misclassified exposure variables. Missingness in these was also likely to be MNAR. For the probabilistic bias analysis and Bayesian analysis, the misclassification parameters were therefore being estimated on different subsets of individuals and with – potentially – different degrees of bias. Similarly, the imputation models for the gold

standard (cotinine) were also applied in these different subsets, again with varying degrees of bias.

Influence of covariates

I did not include covariates in the simulations. The PBA model assumed that misclassification was only differential with respect to the outcome. In reality, some of the covariates (e.g. sex, parental education) could also influence the degree of misclassification in both self-reported and GP-recorded smoking. The multiple imputation models (for the adjusted estimates) took this into account, as the covariates were included in the imputation models. In the Bayesian analysis I attempted to include covariates in the model for true smoking (the exposure model), but these models would not run.

Influence of continuous outcome on misclassification

One important difference between the probabilistic bias analysis and both MI and the Bayesian analysis is that in the PBA, misclassification was only dependent on the measured smoking variable and the binary outcome whereas the MI models included the continuous outcome when imputing the missing true smoking status. Similarly, in the Bayesian analysis, the outcome model (regression of KS4 attainment score on true smoking), the misclassification model, and the exposure model are all part of one overall model and the parameters are jointly estimated; as such, the relationship between the continuous outcome and true smoking feeds into the other parts of the model (i.e. is used when estimating the misclassification parameters).

The advantages and disadvantages of the three methods are summarised in Table 8-9. One of the key advantages of using multiple imputation over probabilistic bias analysis is that MI can simultaneously account for misclassification and missing data, including missing data in the covariates (and outcome, if applicable). The Bayesian analysis could also, in theory, take account of both misclassification and missing data, including missing data in covariates and the outcome of interest. In practice, however, it may be difficult to implement (based on my experience). On this basis, if internal validation data are available, MI appears to be the optimal method

due to its flexibility and because it is relatively easy to implement in standard statistical software packages. If internal validation data were not available, PBA would only be practical in the absence of missing data (or if the complete case analysis would be expected to be unbiased if there were no misclassification). As such, a Bayesian analysis might be the only possible approach in this situation.

Table 8-9: Advantages and disadvantages of the three methods to correct for misclassification

	Advantages	Disadvantages
Probabilistic bias analysis	Can be used when internal validation data are not available.	Can only accommodate missing values in the true exposure – the final analysis will only include only complete cases.
Multiple imputation	MI commands are now available in many statistical software packages. Will take account of misclassification and missing data simultaneously.	Cannot currently be used in the absence of validation data[1].
Bayesian analysis	Can be extended to situations in which internal validation data are not available. As with MI, will address missing data and misclassification simultaneously.	Specialist software is needed; models could take a very long time to run (and in some cases may not run at all). Difficult to extend to missing covariates as a joint model for the missing data must be specified. Models may not always converge or run.

1. Not currently implemented in standard statistical software, although may be possible

I will now focus on the MI results as, for the reasons given above, I believe this to be the most appropriate approach in this instance. When true smoking was MNAR (as likely in the ALSPAC data), the complete case estimates of the effect of smoking for both the binary and continuous outcome were slightly biased towards the null. In general, the extent of the remaining bias (after correcting for misclassification using MI to impute the missing gold standard) if the gold standard were MNAR would be affected by the degree of misclassification. In the ALSPAC data, self-reported smoking

had a similar specificity to GP-recorded smoking (in terms of their relationship with cotinine) but higher sensitivity. In other words, self-reported smoking would be a better "proxy" (the terminology I have used throughout this thesis) for true smoking than GP-recorded smoking. As a result, the corrected estimates of the association between smoking and attainment using self-reported smoking would be expected to be less biased than the corrected estimates using GP-recorded smoking; similarly, the corrected estimates using both smoking variables are likely to reduce bias further (since better, or more, proxies give closer approximations to MAR).

To summarise in terms of the exemplar, the simulations suggest that the naïve analyses using self-reported or GP-recorded smoking would have resulted in estimates of the association between smoking and educational attainment that were biased away from the null. We also know that cotinine was not available on a random subsample of individuals; the same factors that predicted missingness in self-reported smoking predicted missingness in cotinine. In particular, the outcome (educational attainment) was associated with missingness in cotinine and, as discussed in Chapter 7, an interaction between smoking and the binary outcome (did not/did obtain 5+ A*-C grades) with respect to the probability of missingness could not be ruled out. As a result, the complete case estimate of the association between cotinine and educational attainment is also likely to be biased (due to missing data).

The overall MI estimates for cotinine for the binary and continuous outcomes were (OR=) 4.11 (95% CI 3.01, 5.62) and -52 (-60, -44), respectively; these are estimates that take account of both misclassification and missing data. As described above, the results from Chapter 7 indicate that these are likely to be less biased than the complete case estimates but some bias may remain – particularly if, in truth, there is an interaction between cotinine and educational attainment with respect to the probability of missingness. My simulations suggest that this bias would be away from the null. This cannot be ascertained from the data.

In conclusion, my results taken as a whole suggest that daily smokers are (at most) an estimated ~4 times more likely than less than daily or non-smokers to NOT obtain five

or more A* - C grades at GCSE and are likely to obtain capped attainment scores that are (at most) ~50 points lower on average. For any given exemplar, the various results need careful interpretation in the light of the simulations presented here, in order to decide (i) the likely direction of bias and (ii) which results are likely to be the least biased.

Chapter 9 Discussion

Missing data and measurement error – which can also be thought of as a missing data problem – are key sources of potential bias in epidemiological studies. As outlined in the introductory chapters and throughout this thesis, the impact of missing data on estimates of exposure-outcome associations depends on the causes of missingness and the role of these variables (causing missingness) in the analysis model. Similarly, the impact of measurement error on associations also depends on the nature of the error and which variable(s) in the analysis model are measured with error.

Linkage between observational studies and administrative or routine health datasets provides a means of obtaining one or more proxies for study outcomes or exposures. The aim of this thesis was to examine how such proxies obtained from external datasets can be used to investigate and reduce bias due to missing data and measurement error. I investigated this through three exemplar questions and associated simulation studies. In this chapter I summarise my main findings (Section 9.1) and compare these to findings from other studies (Section 9.2). I discuss the strengths and limitations of my work (Section 9.3) and make recommendations in terms of current practice and further research (Sections 9.4 and 9.5, respectively).

9.1 Summary of findings

I have included discussion sections at the end of Chapters 4 to 8 where I have summarised the findings from that chapter and discussed their implications. Therefore, in this chapter I only revisit these briefly rather than discussing them again in detail.

In Chapter 4 I investigated factors associated with ongoing participation in ALSPAC, looking at baseline measures as well as measures obtained from the linked datasets; I also discussed the implications of these findings for the exemplars. The main points from this chapter are summarised in Section 9.1.1. Following this – in Chapters 5 to 7 – I explored the value of linked proxies as a means to understand and reduce bias (in estimates of exposure-outcome associations) due to missing data in a continuous outcome (Chapter 5), a binary outcome (Chapter 6) and a binary exposure (Chapter 7). I summarise these findings in Sections 9.1.2.1 to 9.1.2.3. Chapter 8 explored various methods for correcting for misclassification – in this case in a binary exposure – using both self-reported and linked data, focussing on the situation where a gold (or reference) standard measure was available on a subset of individuals. Again, the focus was on reducing bias in the estimated exposure-outcome association. The findings from this chapter are summarised in Section 9.1.3.

9.1.1 Factors associated with participation in ALSPAC

In Chapter 4 I showed that both child participation (completing questionaires about themselves and attending study clinics) and mother participation (completing questionnaires about themselves or their child) in ALSPAC were associated with a large number of baseline socio-demographic and health-related factors. The key differences between participation by the mother and participation by the child are summarised in the first row of Table 9-1. After adjusting for these baseline factors, linked education variables and variables derived from the GP data were also associated with participation by both the child and mother; again, this is summarised in Table 9-1.

Table 9-1: Summary of key differences in terms of factors associated with participation by the child and the mother

Variable category	Child participation	Mother participation
Baseline factors	Sex (of the child) strong predictor (females more likely to participate)	Sex (of the child) not a predictor Stronger effect of maternal ethnicity and age at first birth
Factors from linked education data	Attainment (at ages 11, 14 and 16) and and school absence at age 15-16 years (higher attainment and lower absence associated with higher participation) Weak association with SEN (individuals with SEN less likely to participate)	Attainment and school absence (higher attainment and lower absence associated with higher participation) No association with SEN
Factors from linked GP data	Strong association with smoking, depression, BMI, GP consultation rates (15-19 years) and prescription rates (15-19 years) (smokers, those with depression and higher BMI less likely to participate; higher consultation and prescription rates associated with higher participation)	Weaker association with depression, consultation rates and prescription rates; other factors similar

In this analysis I considered participation up to age 19 years as this period covered all the measures included in the exemplar questions. I noted in Chapter 4 that many of the measures derived from the linked datasets cannot themselves be causally related to participation in ALSPAC for the whole time period being considered – for example, school absence at age 15-16 years cannot, by definition, directly cause participation in ALSPAC before age 15. As such, there must be other – perhaps unmeasured – factors (associated with the later measures that I considered) that are causes of participation. A consequence of this is that the data for most ALSPAC analyses are unlikely to be MAR conditional on the observed variables.

In terms of the specific exemplars, this analysis provided evidence that IQ (Exemplar 1: breastfeeding and IQ), depression at 18 (Exemplar 2: smoking in pregnancy and offspring depression), and teenage smoking (Exemplar 3: teenage smoking and

educational attainment) were all likely to be missing not at random conditional on the variables included in the relevant analysis model but that inclusion of one or more proxies for these variables as auxiliary variables in multiple imputation models would give a better approximation to MAR. I discuss the implications of this in Section 9.1.2 below when I summarise the findings and conclusions from each of the exemplars.

9.1.2 Use of linked data to address bias due to missing data

The fully adjusted estimates obtained from the different analysis approaches for all three exemplars (Chapters 5 to 7) are given in Table 9-2. In all cases (missing continuous outcome, missing binary outcome and missing binary exposure) where the specified variable was MNAR, inclusion of proxies from linked datasets in multiple imputation resulted in gains in efficiency compared to the complete case analysis; in contrast, using inverse probability weighting often led to a loss in precision compared to the complete case analysis. Further, the results from the simulation studies suggest that – for all three exemplars – the estimates obtained from MI (and the FIML estimates for Exemplar 1) are likely to be the least biased. The simulations also indicate that the estimates obtained using IPW could – in some situations – be more biased than the complete case estimates. This will arise when the missingness model (used to generate the inverse probability weights) is mis-specified. This could happen if, for example, the underlying missingness model is binomial but a logistic model is used to generate the weights, or if the underlying missingness model includes interactions but these are not included in the model used to generate the weights. Below I briefly summarise the findings from each exemplar.

Table 9-2: Estimates of exposure outcome associations from the three exemplars: comparison of approaches to missing data

Exemplar	Outcome[1]	Exposure[1]		Analysis approach			
				Complete case analysis	IPW[2]	MI[2]	FIML[2]
1	IQ (15 years)	Duration of breastfeeding (months):	Never/< 1	0 (ref)[3]	0 (ref)	0 (ref)	0 (ref)
			1 to <3	0.8 (-0.5, 2.1)	1.3 (0.4, 2.3)	1.6 (0.0, 3.1)	1.4 (0.4, 2.3)
			3 to <5	2.6 (1.4, 3.8)	3.0 (2.0, 3.9)	3.0 (1.7, 4.4)	3.0 (2.1, 3.9)
			6+	3.5 (2.5, 4.5)	3.9 (3.1, 4.7)	4.4 (3.2, 5.7)	3.9 (3.1, 4.7)
2	Depression (y/n, 18 yrs)	Smoked in pregnancy (y/n)		1.47 (0.95, 2.27)	1.43 (0.83, 2.44)	1.34 (0.96, 1.89)	N/A
3	Attainment score (16 yrs)	Ever smoked (y/n)		-26 (-30, -23)	-29 (-34, -24)	-31 (-34, -27)	N/A
		Current smoking (daily vs <daily/never)		-51 (-57, -44)	-69 (-85, -53)	-57 (-66, -49)	
		Cotinine (≥9.5 vs <9.5 ng/ml)		-42 (-49, -35)	-57 (-72, -41)	-52 (-60, -44)	
	5+ A*- C grades (no/yes) (16 yrs)	Ever smoked (y/n)		3.32 (2.46, 4.48)	2.74 (1.95, 3.86)	2.56 (2.46, 3.16)	N/A
		Current smoking (daily vs <daily/never)		8.34 (5.25, 13.25)	8.31 (4.99, 13.86)	5.22 (4.01, 6.80)	
		Cotinine (≥9.5 vs <9.5 ng/ml)		4.88 (2.86, 8.31)	5.50 (3.03, 10.01)	4.11 (3.01, 5.62)	

1. ALSPAC-measured variables
2. Using linked variables (education/GP) as auxiliary variables to derive weights (IPW), in MI models, or as extra dependent variables (FIML)
3. Mean difference in IQ points

9.1.2.1 Missing continuous outcome: duration of breastfeeding and IQ

In Chapter 5 I argued that, since there was evidence that IQ was MNAR, the estimates of the impact of duration of breastfeeding on IQ were likely to be biased and that inclusion of linked attainment and other school data as auxiliary variables in MI and FIML models would reduce this bias. The simulations supported this conclusion, suggesting that the inclusion of linked proxies would lead to reductions in bias and gains in efficiency as long as the correlation between the linked proxy and the missing study outcome was at least 0.5. However, the simulations also suggest that the estimates from MI and FIML shown in Table 9-2 are still likely to be under-estimates of the true association between duration of breastfeeding and IQ (under the assumption of no residual confounding).

9.1.2.2 Missing binary outcome: smoking in pregnancy and offspring depression

A complete case logistic regression will give an (asymptotically) unbiased estimate of the odds ratio for exposure as long as there is not a multiplicative interaction between the exposure and outcome with respect to the probability of missingness, with a similar condition required if missingness also depends on covariates (Bartlett et al. 2015). In Chapter 6 I investigated whether this was likely to remain the case if missingness was dependent on an underlying continuous measure (of the outcome). The results of the simulations suggested that the complete case estimate of the exposure odds ratio is likely to be subject to little or no bias if there is no interaction between the exposure and (continuous underlying) outcome with respect to missingness, but there could be substantial bias if an interaction were present. Using a (binary) proxy for the missing binary outcome as an auxiliary variable in MI would be beneficial whether or not an interaction were present – but only if the binary outcome were imputed directly – particularly if this proxy had high sensitivity and specificity. In the absence of an interaction, this would result in gains in efficiency; if an interaction were present, this would also lead to reductions in bias.

In terms of the exemplar, the results suggest – as summarised above – that the MI estimate (including measures of GP-recorded depression as auxiliary variables) is likely to be the least biased. The odds ratio from MI was 1.34 (95% CI 0.96, 1.89) but the simulations suggest that this could still be an over-estimate (if there were an interaction between smoking in pregnancy and offspring depression with respect to the probability of being a complete case). It is difficult to tell whether this is likely to be the case – the results were consistent with there being no interaction between smoking in pregnancy and GP-recorded depression with respect to the probability of being a complete case but this does not rule out an interaction. Finally, evidence from the study by Taylor et al. (Taylor et al. 2017) suggests that this observed association is also likely to be biased due to unmeasured confounding by socio-economic position and other parental factors and, as such, is unlikely to represent a causal link.

9.1.2.3 Missing binary exposure: teenage smoking and educational attainment

In this exemplar I had two different outcome variables – a continuous outcome and a binary outcome. There was evidence that the odds of being a complete case was dependent on both the outcome (educational attainment) and one of the exposures (frequency of smoking). In this scenario, the regression coefficient for smoking (i.e. in the analysis of the continuous attainment score) obtained from the complete case analysis would be expected to be biased. Further, an interaction could not be ruled out (although the data were also consistent with there being no interaction). If an interaction were present, this would mean that the complete case estimate of the odds ratio for smoking (for the binary outcome, not obtaining five or more A*- C grades at GCSE) would also be biased. The simulations suggest that, in the presence of an interaction, the bias for both outcomes (continuous and binary) would be away from the null.

The simulations suggested that the inclusion of one or more proxies for smoking as auxiliary variables in MI would reduce bias and increase efficiency for both a continuous and a binary outcome in situations when missingness was dependent on both the (missing) exposure and the outcome. In the absence of an interaction

between the exposure and outcome with respect to the probability of missingness, the bias (in the estimates of smoking on the outcomes) would be effectively eliminated. In contrast, if an interaction were present, important bias would remain. In all scenarios, the inclusion of two proxies resulted in greater reductions in bias than just using one.

As with Exemplar 2, an interaction between smoking (the exposure) and educational attainment (the outcome) could not be ruled out, although the results were also consistent with there being no interaction. As mentioned previously, the simulations suggest that the MI estimates of the association between smoking and educational attainment would be less biased than the complete case estimates in either situation (with or without an interaction). The implications of this findings for the exemplar are discussed below in Section 9.1.3 because the same exemplar was also used to examine bias due to misclassification.

9.1.3 Correcting for misclassification

In Chapter 8 I revisited Exemplar 3 to examine the impact of misclassification on the association between teenage smoking and educational attainment, both for the binary outcome (not obtaining 5+ A*- C grades) and the continuous outcome (attainment score). The analyses presented in this chapter suggested that misclassification in self-reported and GP-recorded smoking would have resulted in estimates of the impact of smoking on educational attainment that were biased away from the null.

In terms of the methods used to correct for misclassification, I concluded that multiple imputation is the most flexible because it can be used to take account of the misclassification at the same time as taking account of missing data in the other variables in the analysis model. Thus, the MI estimates of the effect of cotinine on educational attainment shown in Table 9-2 above are likely to be the least biased – but, as explained previously, not free from bias. Note that although this analysis could also be done using a fully Bayesian model, in practice this is difficult to implement

because it requires specification of a joint (imputation) model for all the variables. In addition, in my example the Bayesian models proved difficult (in some cases impossible) to run.

To summarise the overall implications for this exemplar, the observed association was large in magnitude, with daily smokers being an estimated 4.11 times more likely to NOT obtain five or more A* to C grades at GCSE and, on average, having attainment scores approximately 50 points lower (just over half a standard deviation) than non-smokers and those smoking less frequently than daily. Some bias due to missing data and misclassification may remain in these estimates and, as such, these might be over-estimates of the association. However, this bias is unlikely to account for the entire association.

Having said this, it is important to remember that, although there are plausible causal mechanisms for such an association, this observed difference could be due to unmeasured confounding. Alternatively, and as discussed in Chapter 2, poor educational attainment has been shown to be associated with increased risk of smoking initiation and lifetime tobacco use. In my analysis I adjusted for earlier attainment (at age 11). Nevertheless, the observed association could still be explained by reverse causation if poor attainment in the early teens resulted in an increased risk of taking up smoking.

Table 9-3 gives what I believe to be the least biased estimate of the association for each exemplar and summarises the conclusions I have made regarding these estimates. (The naïve estimates are included in the table for reference).

Table 9-3: Best estimates of association for each exemplar and summary comments

Exemplar	Outcome	Exposure		Naïve estimate	Best estimate of association	Comments on "best estimate"
1	IQ (15 years)	Duration of breastfeeding (months):	Never/< 1 1 to <3 3 to <5 6+	0 (ref)[3] 0.8 (-0.5, 2.1) 2.6 (1.4, 3.8) 3.5 (2.5, 4.5)	0 (ref) 1.6 (0.0, 3.1) 3.0 (1.7, 4.4) 4.4 (3.2, 5.7)	Likely to be under-estimates of the true association (under the assumption of no residual confounding). Evidence as a whole suggests the association is likely to be causal.
2	Depression (y/n, 18 yrs)	Smoked in pregnancy (y/n)		1.47 (0.95, 2.27)	1.34 (0.96, 1.89)	Likely to be an over-estimate. Evidence as a whole suggests this association is unlikely to be causal – biased due to residual confounding.
3	Attainment score (16 yrs)	Teenage smoking: daily vs <daily/never[1]		-51 (-57, -44)	-52 (-60, -44)	Likely to be over-estimates. Association could be due to reverse causation (or bi-directional); could also be due to residual confounding.
	5+ A*- C grades (no/yes, 16 yrs)			8.34 (5.25, 13.25)	4.11 (3.01, 5.62)	

1. Naïve analysis = based on self-reported smoking; "best estimate" = imputing cotinine using all other measures of smoking, including self-reported and GP-recorded
2. Mean difference in IQ points

9.2 Comparison with other research

Other studies have examined the impact of the inclusion of auxiliary variables in MI models. As in the work presented in this thesis, Collins et al. (Collins et al. 2001) found that, in the scenarios they investigated, the addition of auxiliary variables that were predictors of missingness in the outcome in (linear regression) MI models increased efficiency and reduced bias, even when the correlation between the auxiliary variable and the original outcome variable was relatively low (0.4) – although reductions in bias were relatively small at this level of correlation. However, in their study they only investigated 25% and 50% missing data and only correlations of 0.4 and 0.9. I built on this work by investigating larger proportions of missing data and a wider range of correlation coefficients. I also examined logistic as well as linear regression models.

More recently, Mustillo and Kwon (Mustillo and Kwon 2015) found that the inclusion of auxiliary variables increased efficiency by quite small amounts but did not always reduce bias when data were MNAR. Further, they found that the bias resulting from the data being MNAR was small. This might be explained by the fact that the correlation between their exposure and their outcome in their simulated data was quite high (0.6); further, they only considered up to 30% missing data. In the real dataset in which they simulated missingness, their exposure (not outcome) was MNAR and the analysis was a logistic regression, so the complete case analysis would be expected to be (asymptotically) unbiased. It is not stated how strongly other covariates were related to the exposure variable which they simulated as being MNAR. Again, they only considered 10–30% missing data. As above, I extended this work by examining higher percentages of missing data. I also simulated datasets based on the real ALSPAC data – therefore generating realistic associations between the variables of interest (it would be unusual in an epidemiological study to find a correlation as high as 0.6 between the exposure and outcome).

Two other studies used data available from linked medical records as auxiliary variables in MI models when the outcome variable was MNAR (Hebert et al. 2011,

Wang and Hall 2010); using linked data plus results from simulations, both studies found – as I did – that the inclusion of these auxiliary variables reduced bias but did not eliminate it. However, their work differed from the work presented in this thesis in that – in both of these studies – they were examining bias in the marginal distribution of the outcome variable itself (or the change in the outcome from baseline) rather than in adjusted estimates of the association between an exposure and the missing outcome.

Finally, Ibrahim et al. (Ibrahim et al. 2001) examined the impact of the inclusion of a binary auxiliary variable on maximum likelihood estimates (obtained using an EM algorithm) when a binary outcome was MNAR. They used simulations to vary the correlation between the auxiliary variable and the outcome between 0 and 1. They found that reductions in bias increased as the correlation increased but concluded that the correlation needed to be at least 0.5 for the reduction in bias to be non-trivial. In their dataset, just over 40% of the outcome variable was missing. In this thesis I used MI rather than maximum likelihood methods when I had binary auxiliary variables (proxies), but these findings are similar to mine in that I also found that "better" proxies (i.e. those with higher sensitivity and specificity, or two proxies as opposed to one) led to greater reductions in bias (in the scenarios in which the complete case analysis was biased).

As described in Chapter 2 (Section 2.2.5), previous studies have compared methods to correct for misclassification both in the presence and absence of a gold standard measure. One recent study compared MI to regression calibration and probabilistic bias analysis for a misclassified binary exposure (Livingston et al. 2018). In this study the authors found that PBA performed poorly compared to regression calibration and MI. In addition, they concluded that MI offers a flexible way of correcting for misclassification, particularly as it can also accommodate missingness in the misclassified variable as well as in other covariates (Livingston et al. 2018). Although in my simulated datasets PBA worked well when the outcome was binary, it appeared to result in a small amount of bias in the corrected regression coefficient for the continuous outcome. In the ALSPAC data, the PBA-corrected regression coefficient for the continuous outcome was quite different from that obtained using MI. Further, as I

noted in Chapter 8 and in agreement with the above authors, MI offers the advantage of being able to simultaneously tackle missingness and misclassification.

As noted in Chapter 2, Bartlett and Keogh argued that a fully Bayesian analysis would be preferable to MI in the context of measurement error, partly because of some of the limitations of MI but also because MI cannot be used in the absence of validation data. However, I was not able to identify any studies that compared MI to Bayesian methods empirically. Moreover, the comparisons (of other methods) that I did identify were carried out in the context of completely observed data (apart from in the gold standard measure). One study used Bayesian methods to correct for misclassification and missing data in the absence of a gold standard where a particular variable in two linked datasets was subject to misclassification (He et al. 2014), but this was not compared to other methods. In their study, He et al. assumed the specificity for both measures was 100% but both were subject to under-reporting (i.e. the sensitivities were less than 100%). There was also missing data in both variables; this was assumed to be MAR. They found – as I did – that estimates of exposure-outcome associations corrected for misclassification were, for some variables, very different from the estimates obtained using the misclassified variables.

To my knowledge, methods to correct for misclassification have not been evaluated when the gold standard measure is MNAR, or when the misclassified measure(s) are MNAR. The work I present in this thesis suggests that MI can be used in this context and will reduce bias but not eliminate it. Further work is needed in order to determine whether there are situations in which this no longer applies (either in terms of the level of misclassification or in terms of the extent to which the gold standard is MNAR).

9.3 Strengths and limitations

In this thesis I have extended previous work by examining the impact of the use of auxiliary variables in a wide range of settings. Specifically, I explored missingness in a continuous outcome, a binary outcome and a binary exposure. I investigated whether or not proxies from linked datasets could be used to reduce bias when included as

auxiliary variables in MI, FIML (for the continuous outcome) and IPW models when these study variables were MNAR (conditional on the variables in the analysis model). I also examined the impact of misclassification. By framing this as a missing data problem, I was able to simultaneously address misclassification (missingness in the true exposure) and missing data in other variables in the analysis model.

A key strength of this work is the use of a longitudinal prospective cohort in which there are complex missing data mechanisms. Further, I used the real data examples to build realistic simulation studies in which I investigated many different scenarios, including some in which there was missingness in the proxies. Backing up the analysis with these simulations allowed me to draw conclusions about the relative performance of the different methods in terms of bias.

This work also has limitations. In the simulations, I covered a range of possible scenarios but, for practical reasons, could not consider every possible situation. For example, in the simulations I simulated the probability (for IQ, Exemplar 1, Chapter 5) or log odds (for depression, Exemplar 2, Chapter 6) of missingness in IQ/depression to be linearly related its value. Further, in each exemplar I did not make missingness dependent on unobserved variables. If there were one or more unmeasured factors predictive of missingness then the relative reductions in bias would be lower. If the proxy were strongly associated with these unmeasured factor(s) then use of the proxy could either increase or reduce bias, depending on the magnitude and directions of the relevant associations. As recommended previously (Thoemmes and Rose 2014), careful thought should be given to the likely causal structure between the variables included in the analysis model, the potential proxy variables, and the missingness mechanism in order to identify whether inclusion of a particular proxy is likely to increase or decrease bias.

Any source of linked data is unlikely to have complete population coverage and, in circumstances where linkage to administrative or routine health data requires consent, this may not be obtained for all participants. Further, incomplete information on identifiers may mean that, in some cases, records from the same individual are not linked. This will impact on the potential benefits of obtaining linked

data. Additional issues will also affect the utility of linked data. Firstly, linkage mismatches (whereby a record from one dataset is erroneously linked to a record in the other dataset) can introduce bias (Harron et al. 2017). Secondly, individuals who appear in the linked dataset may differ systematically from those who do not. This applies to the datasets used in this thesis. For example, ALSPAC individuals who attended an independent school at the time of the linkage to the NPD would not have been linked. I demonstrated in Chapter 5 that these individuals were more likely to have higher IQs and more highly educated parents compared to those who were linked. Similarly, linkage to GP data was only possible for those who had been sent fair processing materials and who had not dissented. Males were less likely to have linked GP data, as were children with more educated fathers, those living in private rented accommodation, those whose mother smoking during pregnancy, those who were breastfed for longer, those whose mother was older at the time of their first pregnancy, those whose mother was not married, and those whose family occupational social class was classified as non-manual.

With the continuous outcome, I investigated the impact of missingness in the linked data. I simulated 20% missing linked data and showed that, even when individuals with higher probabilities of having missing outcome data were also more likely to have missing linked data, this had little impact on the results. Thus, even in situations with incomplete coverage of the linked datasets, use of a linked proxy for a continuous outcome that is MNAR is likely to result in gains in efficiency and reductions in bias. Enders (Enders 2008) examined the impact of missingness in auxiliary variables on bias in FIML models, again for a continuous outcome variable, and found that inclusion of auxiliary variables was always beneficial (measured in terms of both bias reduction and efficiency) even when 50% of the values of the auxiliary variable were missing and when the auxiliary variable itself was MNAR, although the reductions in bias were lower than if the auxiliary variable was fully observed. When the auxiliary variable was MNAR he found that imputed parameters relating to the auxiliary variable (for example, the mean value of the auxiliary variable and the regression coefficient from the regression of the outcome on the auxiliary variable) were biased but that this did not generate bias in the parameters of the

analysis model (Enders 2008). I did not simulate missingness – in the sense of having no linked data at all – in the linked binary measures (representing measures from GP data). However, these measures were simulated as providing a misclassified proxy of the missing study outcome or covariate – in other words, an individual's true depression (Exemplar 2, Chapter 6) or smoking (Exemplar 3, Chapter 7) status was assumed to be missing. In both chapters, I demonstrated that having a better proxy (less misclassification) or more than one proxy lead to greater reductions in bias; this suggests that bias reductions would be likely to be smaller if some linked data were missing altogether.

Although cotinine is regarded as a gold standard, it is not a perfect measure of smoking. The key factor that affects its accuracy is the degree to which an individual is exposed to second-hand smoke (Jarvis et al. 2008). I did not assess the impact of having an imperfect reference standard.

9.4 Recommendations for current practice

Missing data inevitably leads to a loss of power. The impact in terms of bias depends on the missing data mechanism. Unfortunately, it is not possible to distinguish between data that are MAR and data that are MNAR using the observed data alone. However, sometimes it is possible to use the observed data to help identify the most likely mechanism – or a set of likely mechanisms. This will help to determine the most appropriate analysis strategy in terms of minimising bias.

Increasingly observational studies in the UK and elsewhere are using linkage to routine and administrative datasets as a means to provide measures on the study participants; indeed, two of the UK's largest cohorts – UK Biobank and the Million Women Study – use linked data as the main source of follow-up data on its participants (Medical Research Council 2014). Here I provide guidelines for dealing with missing data when linked datasets are available. As in this thesis, I focus on the situation in which the question of interest involves estimating the association between an exposure and an outcome. The steps and key questions to be addressed are summarised in Figure 9-1 and outlined on the following pages.

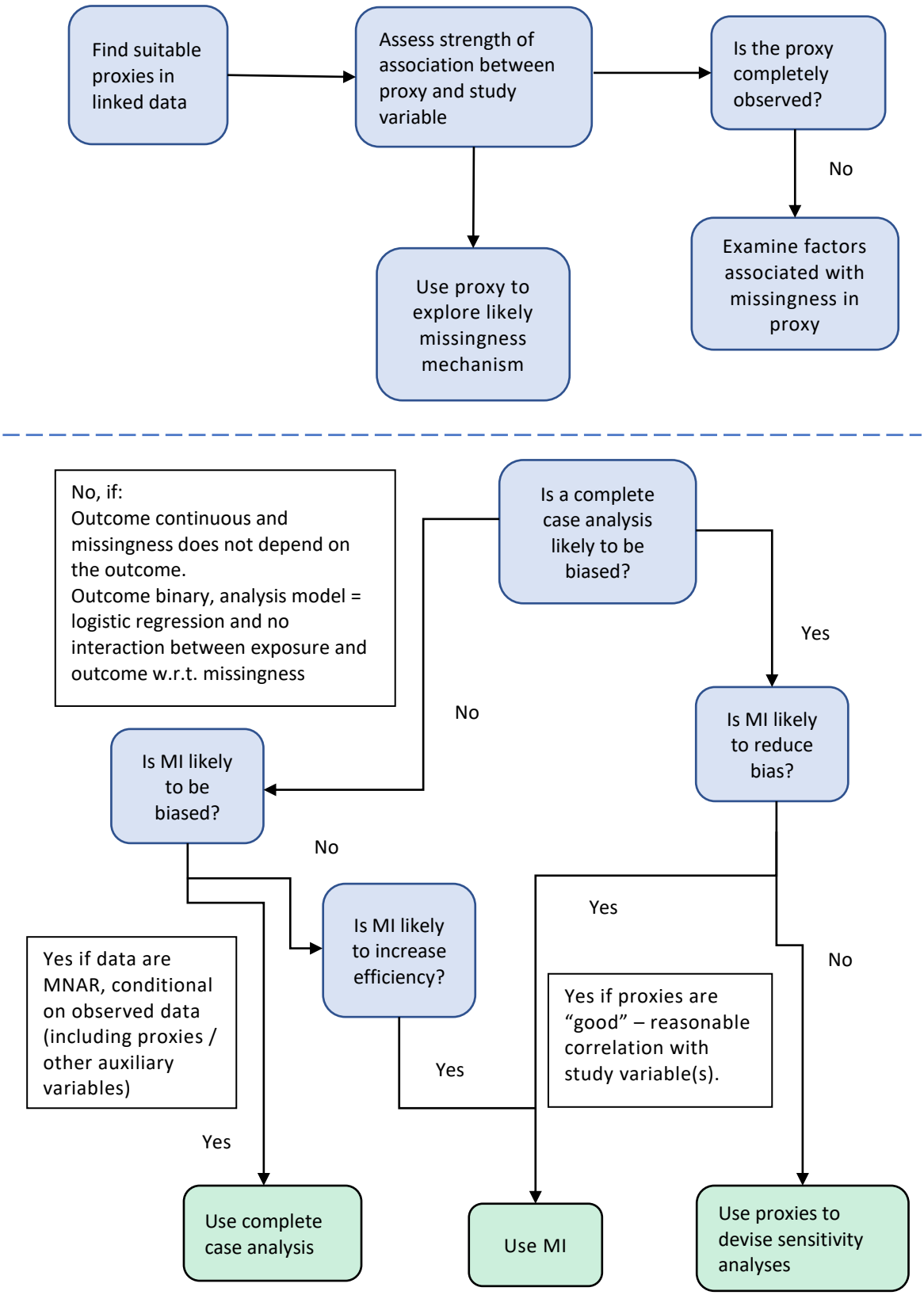


Figure 9-1: Guidelines for addressing missing data when linked datasets are available

Step 1: Finding proxies

These will ideally be variables that are at least moderately correlated with the missing study variable(s). Possibilities include:

- The same variable but from a different source (for example, height, BMI, blood pressure)
- Alternative measures of the same underlying construct (for example, GP-recorded depression, smoking, asthma, etc. as opposed to self-reported/parent-reported measures of these)
- Variables that are strongly correlated with the study variable (for example, attainment as a proxy for IQ)

Step 2: Assessing the strength of association between the proxy and the study variable

For continuous variables this will involve investigating the nature of the association (for example, is it linear?) and calculating the correlation between the proxy and the study variable. For categorical variables, a tabulation may be sufficient. Sensitivity and specificity and/or predictive values can be calculated for binary variables.

Step 3: Use the proxy to explore the likely missingness mechanism

This can be investigated using logistic regression: what factors predict missingness (not being a complete case)? Key questions are:

- Does the outcome (or proxy for the outcome) predict missingness?
- Does the exposure (or proxy for the exposure) predict missingness?
- Are any covariates likely to be MNAR?
- Is there any evidence for an interaction (on the multiplicative scale) between the exposure and outcome with respect to the probability of missingness?

Step 4: What factors predict missingness in the linked proxy/ies (if incomplete)?

Of particular interest is:

- Do different factors predict missingness in the proxy (compared to the study data)?
- Is there any evidence that the proxy could be MNAR?

Step 5: Determining the most appropriate (primary) analysis

The answers from Step 3 will help to decide whether a complete case analysis is likely to be biased. If this is expected to produce an unbiased estimate of the exposure-outcome association then MI will only be preferable if (a) it is also likely to give an unbiased estimate and (b) it is likely to increase efficiency. As summarised in Figure 9-1, this will probably be the case if the data are MAR conditional on the observed data (including any proxies or other auxiliary variables) and the auxiliary variables (including the proxies) are reasonably highly correlated with the missing outcome and/or exposure.

If the complete case analysis is expected to give a biased estimate of the exposure-outcome association then MI is likely to reduce bias if, as above, the proxy/ies (and other auxiliary variables) are reasonably well correlated with the missing outcome and/or exposure. Whether or not MI is expected to reduce bias, the proxy/ies can also be used to devise sensible sensitivity analyses because, as indicated above, they will help to determine a range of plausible missingness mechanisms.

When a gold standard measure is available on a subset of individuals, multiple imputation provides a flexible solution to simultaneously account for both misclassification and missing data in other relevant variables that is straightforward to implement in standard statistical software. As such, I would recommend this approach if internal validation data are available. As this approach treats this as a missing data problem, a similar process to that outlined in Figure 9-1 applies. The "proxies" are now proxies for the true measure and could either be obtained from linked datasets or from the original study data (for example, self-reported measures – as in the exemplar I presented in this thesis). The second step (assessing the strength of association between the proxy and the study variable) is, in this scenario, equivalent to assessing the level of misclassification, since the "study variable" is now the gold standard measure. The questions about whether the complete case analysis is likely to be biased (i.e. using data only from individuals for whom the gold standard measure is available) and whether or not MI will increase efficiency and/or reduce bias remain the same.

Step 6: Carry out simulations

Finally, simulations tailored to the specific dataset and analysis of interest should be used to investigate the likely size and direction of bias under different scenarios. This will help to determine whether important bias remains in the estimate(s) of interest.

The work presented in this thesis also has wider implications for observational epidemiology. In 1998 Egger et al. discussed the pitfalls of combining results from observational studies. In particular, they argued that such studies may give rise to findings that are biased. Thus, combining these (biased) estimates will result in an overall finding that is "very precise but equally spurious" (Egger et al. 1998). This is a very important issue because non-randomised studies – including studies conducted using large datasets of electronic health records – are being increasingly used as a basis for informing clinical practice (Ijaz et al. 2013, Sox and Greenfield 2009). This has led to the development of tools aimed at assessing bias of non-randomised studies when carrying out systematic reviews – for example, the ROBINS-I tool (Sterne et al. 2016).

I found that factors shown to be associated with non-response in epidemiological studies were associated with non-response in ALSPAC. Further, it is plausible that the types of measures that are MNAR will be similar across studies and settings. In addition, it is not unreasonable to assume that the same sorts of measure (self-reported smoking, for example) will be subject to misclassification in different studies and settings. This suggests, as highlighted by Egger et al., that studies investigating a similar epidemiological question are indeed likely to suffer from the same types of bias. Although missing data are increasingly being recognised as an important source of bias, the use and reporting of (correct) methods to address this remains patchy (Perkins et al. 2018). Similarly, a recent review suggested that measurement error is frequently ignored in medical research (Brakenhoff et al. 2018). My findings highlight the importance of assessing the likely extent of these sources of bias and using methods that will minimise it.

9.5 Further work

There are several ways in which the work in this thesis could be developed, some of which arise from the limitations discussed above.

- I assumed a linear relationship between the outcome or exposure of interest and the probability (for IQ: Chapter 5) or log odds (for depression: Chapter 6 and smoking: Chapter 7) of missingness. It would be important to establish whether this has an important impact – either on the resulting bias or on the extent to which this can be reduced by using a linked proxy.

- I investigated a modest amount of missingness in the linked data (where the linked data were MNAR) in Exemplar 1 and found that it had minimal impact. It would be useful to know whether there is a point at which using auxiliary variables from linked datasets would no longer be beneficial, both in terms of the amount of missingness in the linked variable(s) and the extent to which these variables themselves are MNAR (and the mechanism underlying this).

- It would be of interest to investigate more complex scenarios through additional simulations. In particular, it may be important to include a greater number of covariates that are also predictors of missingness. In addition, in my simulations I only generated missing data in the outcome or exposure variable. Further simulations could investigate the impact of having different amounts of missing data in the exposure, the outcome and the covariates.

- My results from Chapter 6 indicated that, if a continuous outcome is MNAR but it is dichotomised and analysed using logistic regression, the complete case estimate of the exposure log odds ratio is unlikely to be biased as long as there is not an interaction between the outcome and exposure with respect to the probability of missingness. However, I did not address this question directly and – again – this may not hold if the missingness mechanism is non-linear.

- Although cotinine is regarded as a gold standard, it is not a perfect measure of smoking. The key factor that affects its accuracy is the degree to which an individual is exposed to second-hand smoke (Jarvis et al. 2008). I did not assess the impact of having an imperfect reference standard.

- My simulations suggested that MI reduced bias due to misclassification even when the gold standard was MNAR. Further simulations could investigate this in greater detail, varying the missingness mechanisms (for the gold standard as well as the proxy/ies), the amount of missing data, and the strength of association between the proxies and the gold standard (i.e. the extent of misclassification).

- I investigated misclassification when internal validation data (using a gold standard measure) were available. Corbin et al. (Corbin et al. 2017) used probabilistic bias analysis and Bayesian methods to correct for misclassification in the absence of validation data. In the context of epidemiological studies like ALSPAC with linkage to routine health data there may be no gold standard or reference test. For example, in my second exemplar depression was measured in ALSPAC by means of the CIS-R, which is based on self-reported symptoms; measures of depression (diagnosis, recorded symptoms, treatment) were also available in the linked GP data. However, none of these measures can be regarded as a gold standard (i.e. all will be subject to misclassification). The work by Corbin et al. could be extended to (i) include more than one misclassified measure of the exposure or outcome and (ii) simultaneously address missing values in the misclassified measures as well as in other variables in the analysis model. Further work could also examine the impact of having an imperfect “gold” standard.

9.6 Overall summary

In Box 9-1 and Box 9-2 I summarise the thesis as a whole, focussing on the question I set out to answer, what this work adds, and what questions remain unanswered.

Box 9-1: Thesis summary: research question and what was already known

Main research question

Can linked health and administrative data be used to reduce bias (in exposure-outcome estimates) due to missing data and misclassification in prospective cohort studies?

Why is this important?

Bias can lead to incorrect conclusions being drawn; this could impact on healthcare and policy decisions. Observational data are being increasingly used as the basis of such decision making.

What was already known

- Missing data results in a loss of power.
- Complete case estimates of exposure-outcome associations will generally be biased if the missing data mechanism depends on the outcome of interest.
- Using auxiliary variables in multiple imputation can reduce bias and increase efficiency if they are reasonably highly correlated with the exposure and/or outcome of interest.
- Misclassification always results in bias.
- When a gold standard is available, MI offers a flexible way of taking account of both missing data and misclassification.

Box 9-2: Thesis summary: main findings and remaining questions

What this work adds

- I have investigated the impact of the use of auxiliary variables (specifically, proxies for missing study variables obtained via linkage to external datasets) in a wide range of conditions using simulation studies based on a real dataset with complex patterns of missing data.
- I demonstrate that bias due to misclassification can be substantial and show that MI can reduce bias due to misclassification even when the gold standard measure is missing not at random.
- I show that linked proxies can reduce bias and improve efficiency in many different scenarios, even if the proxies are themselves incomplete.
- I provide guidance on how to approach missing data and misclassification problems when such proxies are available.

What questions remain unanswered?

- Are there situations in which the use of linked proxies (in MI) becomes detrimental in terms of either bias or efficiency? To answer this, further simulations are needed, investigating (1) different types of missingness mechanism (non-linear) – in the study variables (including any gold standard measure) and the linked proxies, (2) the impact of missing covariates (which are themselves causes of missingness), and (3) the degree of misclassification.
- What is the impact (in terms of bias) of having an imperfect reference standard?
- Which method(s) should be used to address misclassification when there are two or more misclassified measures of the exposure or outcome of interest (all potentially subject to missing data), none of which is a gold standard.

References

Anderson JW, Johnstone BM, Remley DT. Breast-feeding and cognitive development: a meta-analysis. Am J Clin Nutr. 1999;70(4):525-35.

Ashford J, van Lier PA, Timmermans M et al. Prenatal smoking and internalizing and externalizing problems in children studied from childhood to late adolescence. J Am Acad Child Adolesc Psychiatry. 2008;47(7):779-87.

Atkinson MD, Kennedy JI, John A et al. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. BMC Med Inform Decis Mak. 2017;17(1):2.

Azur MJ, Stuart EA, Frangakis C et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40-9.

Banack HR, Stokes A, Fox MP et al. Stratified Probabilistic Bias Analysis for Body Mass Index-related Exposure Misclassification in Postmenopausal Women. Epidemiology. 2018;29(5):604-13.

Barry SJE, Dinnett E, Kean S et al. Are Routinely Collected NHS Administrative Records Suitable for Endpoint Identification in Clinical Trials? Evidence from the West of Scotland Coronary Prevention Study. PLoS One. 2013;8(9).

Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. Am J Epidemiol. 2015.

Bartlett JW, Keogh RH. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Stat Methods Med Res. 2016;27(6):1695-708.

Bartlett JW, Seaman SR, White IR et al. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462-87.

Bebbington P, Brugha T, Coid J et al. Adult psychiatric morbidity in England, 2007: Results of a household survey. 2007.

Bjertness E, Sagatun Å, Green K et al. Response rates and selection problems, with emphasis on mental health variables and DNA sampling, in large population-based, cross-sectional and longitudinal studies of adolescents in Norway. 2010;10(1):602.

Blackwell M, Honaker J, King G. A Unified Approach to Measurement Error and Missing Data: Overview and Applications. Sociological Methods & Research. 2017;46(3):303-41.

Boyd A, Golding J, Macleod J et al. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2012.

Brakenhoff TB, Mitroiu M, Keogh RH et al. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol. 2018;98:89-97.

Bray JW, Zarkin GA, Ringwalt C et al. The relationship between marijuana initiation and dropping out of high school. Health Econ. 2000;9(1):9-18.

Brilleman SL, Salisbury C. Comparing measures of multimorbidity to predict outcomes in primary care: a cross sectional study. Fam Pract. 2012.

Brion MJ, Lawlor DA, Matijasevich A et al. What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. Int J Epidemiol. 2011;40(3):670-80.

Brion MJ, Victora C, Matijasevich A et al. Maternal smoking and child psychological problems: disentangling causal and noncausal effects. Pediatrics. 2010;126(1):e57-65.

Brochu P, Morin L-P, Billette J-M. Opting or Not Opting to Share Income Tax Information with the Census: Does It Affect Research Findings? Canadian Public Policy-Analyse De Politiques. 2014;40(1):67-83.

Brugha TS, Bebbington PE, Jenkins R et al. Cross validation of a general population survey diagnostic interview: a comparison of CIS-R with SCAN ICD-10 diagnostic categories. Psychol Med. 1999;29(05):1029-42.

Busch V, Laninga-Wijnen L, Schrijvers AJP et al. Associations of health behaviors, school performance and psychosocial problems in adolescents in The Netherlands. Health Promot Int. 2017;32(2):280-91.

Carpenter JR, Kenward MG. Multiple Imputation and its Application: Wiley; 2013.

Chu H, Wang Z, Cole SR et al. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. Ann Epidemiol. 2006;16(11):834-41.

Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006;35(4):1074-81.

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330-51.

Collishaw S, Maughan B, Natarajan L et al. Trends in adolescent emotional problems in England: a comparison of two national cohorts twenty years apart. J Child Psychol Psychiatry. 2010;51(8):885-94.

Copeland KT, Checkoway H, McMichael AJ et al. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977;105(5):488-95.

Corbin M, Haslett S, Pearce N et al. A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable. Int J Epidemiol. 2017.

Cornish RP, Boyd A, Van Staa T et al. Socio-economic position and childhood multimorbidity: a study using linkage between the Avon Longitudinal study of parents and children and the general practice research database. International Journal for Equity in Health. 2013;12(1):66.

de Graaf R, Bijl RV, Smit F et al. Psychiatric and sociodemographic predictors of attrition in a longitudinal study: The Netherlands Mental Health Survey and Incidence Study (NEMESIS). Am J Epidemiol. 2000;152(11):1039-47.

Deary IJ, Strand S, Smith P et al. Intelligence and educational achievement. Intelligence. 2007;35(1):13-21.

Deoni SC, Dean DC, 3rd, Piryatinsky I et al. Breastfeeding and early white matter development: A cross-sectional study. Neuroimage. 2013;82:77-86.

Dolan CV, Geels L, Vink JM et al. Testing Causal Effects of Maternal Smoking During Pregnancy on Offspring's Externalizing and Internalizing Behavior. Behav Genet. 2016;46(3):378-88.

Dong Y, Peng C-YJ. Principled missing data methods for researchers. Springerplus. 2013;2.

Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol. 1990;132(4):746-8.

Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989;8(5):551-61.

Edwards JK, Cole SR, Troester MA et al. Accounting for Misclassified Outcomes in Binary Regression Models Using Multiple Imputation With Internal Validation Data. Am J Epidemiol. 2013;177(9):904-12.

Eekhout I, de Boer RM, Twisk JW et al. Missing data: a systematic review of how they are reported and handled. Epidemiology. 2012;23(5):729-32.

Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. BMJ. 1998;316(7125):140-4.

Ekblad M, Gissler M, Lehtonen L et al. Prenatal smoking exposure and the risk of psychiatric morbidity into young adulthood. Arch Gen Psychiatry. 2010;67(8):841-9.

Enders CK. A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. Structural Equation Modeling-a Multidisciplinary Journal. 2008;15(3):434-48.

Enders CK. Applied Missing Data Analysis. New York: The Guildford Press; 2010.

Faris PD, Ghali WA, Brant R et al. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidemiol. 2002;55(2):184-91.

Farrell P, Fuchs VR. Schooling and health: the cigarette connection. J Health Econ. 1982;1(3):217-30.

Ferrari AJ, Charlson FJ, Norman RE et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med. 2013;10(11):e1001547.

Ferrie JE, Kivimaki M, Singh-Manoux A et al. Non-response to baseline, non-response to follow-up and mortality in the Whitehall II cohort. Int J Epidemiol. 2009;38(3):831-37.

Ford DV, Jones KH, Verplancke JP et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res. 2009;9:157.

Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol. 2005;34(6):1370-6.

Fraser A, Macdonald-Wallis C, Tilling K et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97-110.

Freedman LS, Midthune D, Carroll RJ et al. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. 2008;27(25):5195-216.

Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Current epidemiology reports. 2014;1(4):175-85.

Gage SH, Bowden J, Davey Smith G et al. Investigating causality in associations between education and smoking: a two-sample Mendelian randomization study. Int J Epidemiol. 2018;47(4):1131-40.

Galea S, Tracy M. Participation Rates in Epidemiologic Studies. Ann Epidemiol. 2007;17(9):643-53.

Galimard JE, Chevret S, Protopopescu C et al. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Stat Med. 2016;35(17):2907-20.

Galvan A, Poldrack RA, Baker CM et al. Neural correlates of response inhibition and cigarette smoking in late adolescence. Neuropsychopharmacology. 2011;36(5):970-8.

Gibbs BG, Forste R. Breastfeeding, Parenting, and Early Cognitive Development. The Journal of pediatrics. 2014;164(3):487-93.

Gilbert R, Martin RM, Donovan J et al. Misclassification of outcome in case–control studies: Methods for sensitivity analysis. Stat Methods Med Res. 2016;25(5):2377-93.

Gilman SE, Abrams DB, Buka SL. Socioeconomic status over the life course and stages of cigarette use: initiation, regular use, and cessation. J Epidemiol Community Health. 2003;57(10):802-8.

Gilman SE, Martin LT, Abrams DB et al. Educational attainment and cigarette smoking: a causal association?(). Int J Epidemiol. 2008;37(3):615-24.

Glendinning A, Hendry L, Shucksmith J. Lifestyle, health and social class in adolescence. Soc Sci Med. 1995;41(2):235-48.

Graham JW. Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal. 2003;10(1):80-100.

Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol. 1980;112(4):564-9.

Greenland S. The Impact of Prior Distributions for Uncontrolled Confounding and Response Bias. Journal of the American Statistical Association. 2003;98(461):47-54.

Greenland S. Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. Int J Epidemiol. 2009;38(6):1662-73.

Greenland S, Kleinbaum DG. Correcting for Misclassification in Two-Way Tables and Matched-Pair Studies. Int J Epidemiol. 1983;12(1):93-97.

Hankin BL. Adolescent depression: description, causes, and interventions. Epilepsy Behav. 2006;8(1):102-14.

Harald K, Salomaa V, Jousilahti P et al. Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972-92. J Epidemiol Community Health. 2007;61(5):449-54.

Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. 2001.

Harron KL, Doidge JC, Knight HE et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699-710.

He YL, Landrum MB, Zaslavsky AM. Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: a multiple imputation approach. Stat Med. 2014;33(21):3710-24.

Hebert PL, Taylor LT, Wang JJ et al. Methods for Using Data Abstracted from Medical Charts to Impute Longitudinal Missing Data in a Clinical Trial. Value Health. 2011;14(8):1085-91.

Heilbrun LK, Nomura A, Stemmermann GN. The effects of non-response in a prospective study of cancer: 15-year follow-up. Int J Epidemiol. 1991;20(2):328-38.

Hennekens CH, Buring JE. Chapter 11: Analysis of Epidemiologic Studies: Evaluating the Role of Bias. In: S.L. M, editor. Epidemiology in Medicine. Boston: Little, Brown and Company; 1987.

Hoff PD. A First Course in Bayesian Statistical Methods: Springer; 2009.

Horta BL, Loret de Mola C, Victora CG. Breastfeeding and intelligence: a systematic review and meta-analysis. Acta Paediatr. 2015;104(467):14-9.

Horta BLV, C.G. Long-term effects of breastfeeding: a systematic review. WHO; 2013.

Hosmer DWL, S. . Applied logistic regression. New York: Wiley; 1989.

Hughes RA, White IR, Seaman SR et al. Joint modelling rationale for chained equations. BMC Med Res Methodol. 2014;14(1):28.

Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. BMJ. 2010;340.

Ibrahim JG, Lipsitz SR, Horton N. Using Auxiliary Data for Parameter Estimation with Non-Ignorably Missing Outcomes. Journal of the Royal Statistical Society Series C (Applied Statistics). 2001;50(3):361-73.

Ijaz SI, Mischke C, Ruotsalainen J et al. 179 The use of non-randomised studies in systematic reviews of intervention effectiveness: A content analysis of cochrane systematic reviews. 2013;70(Suppl 1):A60-A60.

Indredavik MS, Brubakk AM, Romundstad P et al. Prenatal smoking exposure and psychiatric symptoms in adolescence. Acta Paediatr. 2007;96(3):377-82.

Isaacs EB, Fischl BR, Quinn BT et al. Impact of breast milk on intelligence quotient, brain size, and white matter development. Pediatr Res. 2010;67(4):357-62.

Jacobson SW, Carter RC, Jacobson JL. Breastfeeding as a proxy for benefits of parenting skills for later reading readiness and cognitive competence. J Pediatr. 2014;164(3):440-2.

Jarvis MJ, Fidler J, Mindell J et al. Assessing smoking status in children, adolescents and adults: cotinine cut-points revisited. Addiction. 2008;103(9):1553-61.

Jarvis MJ, Tunstall-Pedoe H, Feyerabend C et al. Comparison of tests used to distinguish smokers from nonsmokers. Am J Public Health. 1987;77(11):1435-8.

John A, McGregor J, Fone D et al. Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. BMC Med Inform Decis Mak. 2016;16(1):35.

Joseph L, Gyorkos TW, Coupal L. Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard. Am J Epidemiol. 1995;141(3):263-72.

Jurek AM, Greenland S, Maldonado G et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. Int J Epidemiol. 2005;34(3):680-7.

Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res. 2007;16(3):199-218.

Kline RB. Principles and practice of structural equation modeling. 3rd edition ed. New York: The Guilford Press; 2011.

Knudsen AK, Hotopf M, Skogen JC et al. The Health Status of Nonparticipants in a Population-based Health Study The Hordaland Health Study. Am J Epidemiol. 2010;172(11):1306-14.

Koivusilta L, Arja R, Andres V. Health behaviours and health in adolescence as predictors of educational level in adulthood: a follow-up study from Finland. Soc Sci Med. 2003;57(4):577-93.

Koivusilta LK, West P, Saaristo VM et al. From childhood socio-economic position to adult educational level - do health behaviours in adolescence matter? A longitudinal study. BMC Public Health. 2013;13:711.

Kramer MS, Aboud F, Mironova E et al. Breastfeeding and child cognitive development: new evidence from a large randomized trial. Arch Gen Psychiatry. 2008;65(5):578-84.

Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. Epidemiology. 1992;3(3):210-5.

Kristensen P, Irgens LM. Maternal reproductive history: a registry based comparison of previous pregnancy data derived from maternal recall and data obtained during the actual pregnancy. Acta Obstet Gynecol Scand. 2000;79(6):471-77.

Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. 1 ed: Springer-Verlag; 2009.

Lavigne JV, Hopkins J, Gouze KR et al. Is Smoking During Pregnancy a Risk Factor for Psychopathology in Young Children? A Methodological Caveat and Report on Preschoolers. J Pediatr Psychol. 2011;36(1):10-24.

Lewis G, Araya R. Classification, disability and the public health agenda: Depression and public health. Br Med Bull. 2001;57(1):3-15.

Lewis G, Pelosi AJ, Araya R et al. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. Psychol Med. 1992;22(2):465-86.

Little R. Regression With Missing X's: A Review. Journal of the American Statistical Association. 1992;87(420):1227-37.

Livingston MDIIIMPH, Cannell B, Muller K et al. Comparing methods of misclassification correction for studies of adolescent alcohol use. The American journal of drug and alcohol abuse. 2018;44(2):160-66.

Lorant V, Demarest S, Miermans PJ et al. Survey error in measuring socio-economic risk factors of health status: a comparison of a survey and a census. Int J Epidemiol. 2007;36(6):1292-99.

Lucas A, Morley R, Cole TJ et al. Breast-Milk and Subsequent Intelligence Quotient in Children Born Preterm. Lancet. 1992;339(8788):261-64.

Lundberg I, Thakker KD, Hallstrom T et al. Determinants of non-participation, and the effects of non-participation on potential cause-effect relationships, in the PART study on mental disorders. Soc Psychiatry Psychiatr Epidemiol. 2005;40(6):475-83.

Lyles RH, Zhang F, Drews-Botsch C. Combining internal and external validation data to correct for exposure misclassification: a case study. Epidemiology. 2007;18(3):321-8.

Lynskey MT, Coffey C, Degenhardt L et al. A longitudinal study of the effects of adolescent cannabis use on high school completion. Addiction. 2003;98(5):685-92.

MacLehose RF, Olshan AF, Herring AH et al. Bayesian Methods for Correcting Misclassification: An Example from Birth Defects Epidemiology. Epidemiology. 2009;20(1):27-35.

Macleod J, Smith GD, Heslop P et al. Psychological stress and cardiovascular disease: empirical demonstration of bias in a prospective observational study of Scottish men. BMJ. 2002;324(7348):1247.

Marmot M. Fair society, healthy lives: the Marmot Review: strategic review of health inequalities in England post-2010. 2010.

Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. J Clin Epidemiol. 1990;43(9):941-7.

Marshall RJ, Zhang Z, Broad JB et al. Agreement between ethnicity recorded in two New Zealand health databases: effects of discordance on cardiovascular outcome measures (PREDICT CVD3). Aust N Z J Public Health. 2007;31(3):211-16.

Martikainen P, Laaksonen M, Piha K et al. Does survey non-response bias the association between occupational social class and health? Scandinavian Journal of Public Health. 2007;35(2):212-15.

Martin J, Tilling K, Hubbard L et al. Association of Genetic Risk for Schizophrenia With Nonparticipation Over Time in a Population-Based Cohort Study. Am J Epidemiol. 2016;183(12):1149-58.

McInturff P, Johnson WO, Cowling D et al. Modelling risk when binary outcomes are subject to error. Stat Med. 2004;23(7):1095-109.

Medical Research Council. Maximising the value of UK population cohorts: MRC Strategic Review of the Largest UK Population Cohort Studies. UK: Medical Research Council; 2014.

Menezes AMB, Murray J, László M et al. Happiness and Depression in Adolescence after Maternal Smoking during Pregnancy: Birth Cohort Study. PLoS One. 2013;8(11):e80370.

Meng X-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. 1994:538-58.

Monshouwer K, Huizink AC, Harakeh Z et al. Prenatal smoking exposure and the risk of behavioral problems and substance use in adolescence: the TRAILS study. Eur Addict Res. 2011;17(6):342-50.

Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. ArXiv e-prints [Internet]. 2017 December 01, 2017. Available from: https://ui.adsabs.harvard.edu/#abs/2017arXiv171203198M.

Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. 2014;14(1):75.

Moylan S, Gustavson K, Øverland S et al. The impact of maternal smoking during pregnancy on depressive and anxiety behaviors in children: the Norwegian Mother and Child Cohort Study. BMC Med. 2015;13(1):1-12.

Moylan S, Jacka FN, Pasco JA et al. How cigarette smoking may increase the risk of anxiety symptoms and anxiety disorders: a critical review of biological pathways. Brain and Behavior. 2013;3(3):302-26.

Mukherjee M, Gupta R, Farr A et al. Estimating the incidence, prevalence and true cost of asthma in the UK: secondary analysis of national stand-alone and linked databases in England, Northern Ireland, Scotland and Wales—a study protocol. BMJ Open. 2014;4(11).

Mustillo S, Kwon S. Auxiliary Variables in Multiple Imputation When Data Are Missing Not at Random. The Journal of Mathematical Sociology. 2015;39(2):73-91.

Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging Themes in Epidemiology. 2017;14.

NHS Digital. Statistics on Smoking - England , 2018. NHS Digital; 2018.

Nilsen RM, Vollset SE, Gjessing HK et al. Self-selection and bias in a large prospective pregnancy cohort in Norway. Paediatr Perinat Epidemiol. 2009;23(6):597-608.

Nohr EA, Frydenberg M, Henriksen TB et al. Does Low Participation in Cohort Studies Induce Bias? Epidemiology. 2006;17(4):413-18.

Nummela O, Sulander T, Helakorpi S et al. Register-based data indicated nonparticipation bias in a health study among aging people. J Clin Epidemiol. 2011;64(12):1418-25.

Orlebeke JF, Knol DL, Verhulst FC. Child behavior problems increased by maternal smoking during pregnancy. Arch Environ Health. 1999;54(1):15-9.

Osler M, Kriegbaum M, Christensen U et al. Loss to follow up did not bias associations between early life factors and adult depression. J Clin Epidemiol. 2008;61(9):958-63.

Patalay P, Fitzsimons E. Mental ill-health among children of the new century: trends across childhood with a focus on age 14. . London: Centre for Longitudinal Studies; 2017.

Pennanen M, Haukkala A, de Vries H et al. Longitudinal study of relations between school achievement and smoking behavior among secondary school students in Finland: results of the ESFA study. Subst Use Misuse. 2011;46(5):569-79.

Perkins NJ, Cole SR, Harel O et al. Principled Approaches to Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):568-75.

Randall DA, Lujic S, Leyland AH et al. Statistical methods to enhance reporting of Aboriginal Australians in routine hospital records using data linkage affect estimates of health disparities. Aust N Z J Public Health. 2013;37(5):442-49.

Ressler KJ, Nemeroff CB. Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. Depress Anxiety. 2000;12 Suppl 1:2-19.

Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. 1993;12(18):1703-22.

Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. Stat Med. 1997;16(1-3):39-56.

Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol. 1999;28(5):964-74.

Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall/CRC Press; 1997.

Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.

Schafer JL. Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. Statistica Neerlandica. 2003;57(1):19-35.

Schomaker M, Gsponer T, Estill J et al. Non-ignorable loss to follow-up: correcting mortality estimates based on additional outcome ascertainment. Stat Med. 2014;33(1):129-42.

Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.

Smoking in Pregnancy Challenge Group. Smoking in Pregnancy Challenge Group: Review of the Challenge 2018. 2018.

Sogaard AJ, Selmer R, Bjertness E et al. The Oslo Health Study: The impact of self-selection in a large, population-based survey. Int J Equity Health. 2004;3(1):3.

Sox HC, Greenfield S. Comparative Effectiveness Research: A Report From the Institute of Medicine. Ann Intern Med. 2009;151(3):203-05.

Sterne JA, Hernán MA, Reeves BC et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. 2016;355:i4919.

Sterne JA, White IR, Carlin JB et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

Stiby AI, Hickman M, Munafò MR et al. Adolescent cannabis and tobacco use and educational outcomes at age 16: birth cohort study. Addiction. 2015;110(4):658-68.

Sun B, Perkins NJ, Cole SR et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data. Am J Epidemiol. 2018;187(3):585-91.

Sun B, Tchetgen Tchetgen EJ. On Inverse Probability Weighting for Nonmonotone Missing at Random Data. Journal of the American Statistical Association. 2018;113(521):369-79.

Taylor AE, Carslake D, de Mola CL et al. Maternal Smoking in Pregnancy and Offspring Depression: a cross cohort and negative control study. Sci Rep. 2017;7(1):12579.

Taylor AE, Jones HJ, Sallis H et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2018;47(4):1207-16.

Thoemmes F, Rose N. A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems. Multivariate Behavioral Research. 2014;49(5):443-59.

Tilling K, Williamson EJ, Spratt M et al. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. J Clin Epidemiol. 2016;80:107-15.

Tin ST, Woodward A, Ameratunga S. Estimating bias from loss to follow-up in a prospective cohort study of bicycle crash injuries. Inj Prev. 2014;20(5):8.

Valle D, Lima JMT, Millar J et al. Bias in logistic regression due to imperfect diagnostic test results and practical correction approaches. 2015;14(1):434.

van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219-42.

van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18(6):681-94.

van Walraven C. Bootstrap imputation with a disease probability model minimized bias from misclassification due to administrative database codes. J Clin Epidemiol. 2017;84:114-20.

van Walraven C. A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. Int J Epidemiol. 2017:dyx253-dyx53.

Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2010;6(1):37-48.

Victora CG, Bahl R, Barros AJD et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet. 2016;387(10017):475-90.

Walfisch A, Sermer C, Cressman A et al. Breast milk and cognitive development—the role of confounders: a systematic review. BMJ Open. 2013;3(8).

Wang C, Hall CB. Correction of Bias from Non-random Missing Longitudinal Data Using Auxiliary Information. Stat Med. 2010;29(6):671-9.

Wang Y, Hunt K, Nazareth I et al. Do men consult less than women? An analysis of routinely collected UK general practice data. BMJ Open. 2013;3(8).

Webb P, Bain C. Chapter 7: All that glitters is not gold. Essential Epidemiology. Second ed. Cambridge, UK: Cambridge University Press; 2011.

Wechsler D. Wechsler Abbreviated Scale of Intelligence. New York: The Psychological Corporation: Harcourt Brace & Company; 1999.

White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920-31.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.

Wigertz A, Lonn S, Hall P et al. Non-participant characteristics and the association between socioeconomic factors and brain tumour risk. J Epidemiol Community Health. 2010;64(8):736-43.

Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. Journal of the American Statistical Association. 2005;100(472):1123-32.

Appendix A: Additional information on variables

Section 1: Questions used to define duration of breastfeeding, smoking in pregnancy and teenage smoking

1. Questions used to derive duration of breastfeeding.

At 4 weeks:

B1. How have you fed your baby since she was born? Please indicate for each of the times given.

		Breast only	Bottle only	Breast and bottle	Other (please describe below)
a)	First 24 hours	1	2	3	4
b)	Rest of 1st week	1	2	3	4
c)	2nd week	1	2	3	4
d)	3rd week	1	2	3	4
e)	4th week	1	2	3	4

……..

B4. a) How is your baby being fed at the moment?

breast	1
bottle	2
breast and bottle	3
other (please describe)	4

……

B7. How often is your baby fed in the following ways:

		Always	Often	Some-times	Never	Don't know
a)	lying propped up (e.g. with a pillow)	1	2	3	4	9
b)	baby lying down with bottle held by you or someone else	1	2	3	4	9
c)	fed with a bottle while held in someone's arms	1	2	3	4	9
d)	Breastfed	1	2	3	4	9

At 6 months:

C2. Did you breast feed?

Yes, I am still breast feeding	1	How many times a day?		
Yes, I breast fed but have now stopped	2	How old was the baby when you stopped?	months	weeks
I never breast fed	3			

At 15 months:

D2. Was he breast fed?

Yes, he is still being breast fed	1 →	How many times a day?	times
Yes, was breast fed but now stopped	2 →	How old was he when breastfeeding stopped?	months
He was never breast fed	3	(Put 00 if less than 1 month)	

2. Questions about smoking in pregnancy

At 18 weeks gestation:

G3. f) Did you smoke regularly at any of the following times in the last 9 months?

	Before pregnancy	First 3 months of pregnancy	Last 2 weeks
No	1	1	1
Yes, cigarettes	2	2	2
Yes, cigar	3	3	3
Yes, pipe	4	4	4
Yes, other (please describe)	5	5	5

……

At 32 weeks gestation:

E3. How many cigarettes per day are you yourself smoking at the moment? cigarettes

At 8 weeks (after birth):

B4. Did you smoke regularly in the last 2 months of preg nancy and since having the baby?

	(a) Last 2 months of pregnancy		(b) Since having the baby	
	Yes	No	Yes	No
i) cigarettes	1	2	1	2
ii) cigar	1	2	1	2
iii) pipe	1	2	1	2
iv) other	1	2	1	2

3. Questions about teenage smoking (age 15)

Have you ever tried a cigarette (including roll-ups), even a puff?
Have you smoked any cigarettes in the past 30 days?
Do you smoke every day?
How many cigarettes do you smoke per day, on average?
Do you smoke every week?
How many cigarettes do you smoke per week, on average?

Section 2: Read code lists

Table 1: Read codes used to define a diagnosis of depression, John et al.

Read code	Description
Eu32.	[X]Depressive episode
Eu320	[X]Mild depressive episode
Eu321	[X]Moderate depressive episode
Eu322	[X]Severe depressive episode without psychotic symptoms
Eu324	[X]Mild depression
Eu32y	[X]Other depressive episodes
Eu32z	[X]Depressive episode, unspecified
Eu33.	[X]Recurrent depressive disorder
Eu330	[X]Recurrent depressive disorder, current episode mild
Eu331	[X]Recurrent depressive disorder, current episode moderate
Eu332	[X]Recurrent depressive disorder, current episode severe without psychotic symptoms
Eu334	[X]Recurrent depressive disorder, currently in remission
Eu33y	[X]Other recurrent depressive disorders
Eu33z	[X]Recurrent depressive disorder, unspecified
Eu341	[X]Dysthymia
E118.	Seasonal affective disorder
E135.	Agitated depression
E2B..	Depressive disorder NEC
E2B1.	Chronic depression
E291.	Prolonged depressive reaction
E204.	Neurotic depression reactive type
E2B0.	Postviral depression
E112.	Single major depressive episode
E1120	Single major depressive episode, unspecified
E1121	Single major depressive episode, mild
E1122	Single major depressive episode, moderate
E1123	Single major depressive episode, severe, without psychosis
E1125	Single major depressive episode, partial or unspecied remission
E1126	Single major depressive episode, in full remission
E112z	Single major depressive episode NOS
E113.	Recurrent major depressive episode
E1130	Recurrent major depressive episodes, unspecified
E1131	Recurrent major depressive episodes, mild
E1132	Recurrent major depressive episodes, moderate
E1133	Recurrent major depressive episodes, severe, no psychosis
E1135	Recurrent major depressive episodes, partial/unspecified remission
E1136	Recurrent major depressive episodes, in full remission
E1137	Recurrent depression
E113z	Recurrent major depressive episode NOS
E2003	Anxiety with depression
Eu412	[X]Mixed anxiety and depressive disorder

Table 2: Read codes used to define symptoms of depression, John et al.

Read code	Description
1B17.	Depressed
1B1U.	Symptoms of depression
1BQ..	Loss of capacity for enjoyment
1BT..	Depressed mood
1BU..	Loss of hope for the future
2257.	O/E – depressed

Table 3: Read codes used to define treatment for depression, John et al.

Read code	Drug name
d71..	Amitriptyline hydrochloride
d72..	Butriptyline - discontinued
d73..	Clomipramine hydrochloride
d74..	Desipramine hydrochloride
d75..	Dosulepin Hydrochloride
d76..	Doxepin
d77..	Imipramine hydrochloride
d78..	Iprindole
d79..	Lofepramine
d7a..	Maprotiline hydrochloride
d7b..	Mianserin hydrochloride
d7c..	Nortriptyline
d7d..	Protriptyline hydrochloride
d7e..	Trazadone hydrochloride
d7f..	Trimipramine
d7g..	Viloxazine hydrochloride
d7h..	Amoxapine
d81..	Phenelzine
d83..	Isocarboxazid
d84..	Tranylcypromine
d85..	Moclobemide
d91..	Compound Antidepressants A-Z
da1..	Flupentixol [Antidepressant]
da2..	Tryptophan
da3..	Fluvoxamine Maleate
da4..	Fluoxetine hydrochloride
da5..	Sertraline hydrochloride
da6..	Paroxetine hydrochloride
da7..	Venlafaxine
da9..	Citalopram
daA..	Reboxetine
daB..	Mirtazapine
daC..	Escitalopram
daD..	Agomelatine
gde..	Duloxetine

Table 4: Read codes (signs and symptoms) for smoking status, Atkinson et al.

Read code	Description	Smoking status
137..	Tobacco consumption	S1
1371.	Never smoked tobacco	N
1372.	Trivial smoker <1 cig/day	S
1373.	Light smoker 1-9 cigs/day	S
1374.	Moderate smoker 10-19 cigs/day	S
1375.	Heavy smoker 20-39 cigs/day	S
1376.	Very heavy smoker 40+ cigs/day	S
1377.	Ex-trivial smoker (<1/day)	E
1378.	Ex-light smoker (1-9/day)	E
1379.	Ex-moderate smoker (10-19/day)	E
137A.	Ex-heavy smoker (20-39/day)	E
137a.	Pipe tobacco consumption	S
137B.	Ex-very heavy smoker (40+/day)	E
137b.	Ready to stop smoking	S
137C.	Keeps trying to stop smoking	S
137c.	Thinking about stopping smoking	S
137D.	Admitted tobacco consumption untrue	S
137d.	Not interested in stopping smoking	S
137e.	Smoking restarted	S
137E.	Tobacco consumption unknown	S1
137F.	Ex-smoker amount unknown	E
137f.	Reason for restarting smoking	S
137g.	Cigarette pack years	S1
137G.	Trying to give up smoking	S
137h.	Minutes from waking to first tobacco	S
137H.	Pipe smoker	S
137i.	Ex-tobacco chewer	E
137J.	Cigar smoker	S
137j.	Ex cigarette smoker	E
137K.	Stopped smoking	E
137K0	Recently stopped smoking	E
137L.	Current non-smoker	E
137l.	Ex roll up cigarette smoker	E
137m.	Failed attempt to stop smoking	S
137M.	Rolls own cigarettes	S
137N.	Ex pipe smoker	E
137O.	Ex cigar smoker	E
137P.	Cigarette smoker	S
137Q.	Smoking started	S
137R.	Current smoker	S
137S.	Ex smoker	E
137T.	Date ceased smoking	E
137V.	Smoking reduced	S
137X.	Cigarette consumption	S1
137Y.	Cigar consumption	S1
137Z	Tobacco consumption NOS	S1
13p..	Smoking cessation milestones	S
13p0.	Negotiated date for cessation of smoking	S
13p4.	Smoking free weeks	E
13p5.	Smoking cessation programme start date	S
13p8.	Lost to smoking cessation follow up	S

1. For these codes, the value associated with the code had to be greater than zero.

Table 5: Read codes (other) for smoking status, Atkinson et al.

Read code	Description	Smoking status
38DH.	Fagerstron test for nicotine dependence	S
6791.	Health ed - smoking	S
67910	Health ed - parental smoking	S
67A3.	Pregnancy smoking advice	S
67H1.	Lifestyle advice regarding smoking	S
67H6.	Brief cessation for smoking cessation	S
745H%	Smoking cessation therapy	S
8B2B.	Nicotine replacement therapy	S
8B3f.	Nicotine replacement therapy provided free	S
8B3Y.	Over the counter nicotine replacement therapy	S
8BP3.	NRT provided by community pharmacist	S
8CAg.	Smoking cessation therapy provided by community pharmacist	S
8CAL.	Smoking cessation advice	S
8CdB.	Stop smoking service opportunity signposted	S
8H7i.	Referral to smoking cessation advisor	S
8HBM.	Stop smoking face to face follow up	S
8HkQ.	Referral to NHS stop smoking service	S
8HTK.	Referral to stop smoking clinic	S
8I2I.	Nicotine replacement therapy contraindicated	S
8I2J.	Bupropion contraindicated	S
8I39.	Nicotine replacement therapy refused	S
8I3M.	Bupropion refused	S
8I6H.	Smoking review not indicated	S
8IAj.	Smoking cessation advice declined	S
8IEK.	Smoking cessation programme declined	S
8IEM.	Smoking cessation drug therapy declined	S
9hG..	Exception reporting: smoking quality indicators	S
9hG0.	Excepted from smoking quality indicators: patient unsuitable	S
9hG1.	Excepted from smoking quality indicators: informed dissent	S
9kc..	Smoking cessation - enhanced services admin	S
9kc0.	Smoking cessation monitoring template completed	S
9km..	Ex-smoker annual review	E
9kn..	Non-smoker annual review	N
9ko..	Current smoker annual review	S
9N2k.	Seen by smoking cessation advisor	S
9N4M.	DNA smoking cessation clinic	S
9Ndg.	Declined consent for follow up by smoking cessation team	S
9NdV.	Consent given for follow up after smoking cessation intervention	S
9NdW.	Consent given for smoking cessation data sharing	S
9NdY.	Declined consent for follow up evaluation after smoking cessation intervention	S
9NdZ.	Declined consent for smoking cessation data sharing	S
9NS02	Referral for smoking cessation service offered	S
9OO%	Attends stop smoking monitor admin	S
E023.	Nicotine withdrawal	S
E251%	Tobacco dependence	S
J0364	Tobacco deposit on teeth	S
SMC..	Toxic effect of tobacco and nicotine	S
TJHy2	Adverse reaction to nicotine	S
U6099	[X] Bupropion causing adverse effects in therapeutic use	S
ZV4K0	[V] Tobacco use	S
ZV6D8	[V] Tobacco abuse counselling	S

Table 6: Read codes (drugs) for smoking status, Atkinson et al.

Read code	Drug (various products)
du3%	Nicotine
du6%	Bupropion
du7%	Nicotine
du8%	Varenicline
du9%	Nicotine

Appendix B: Additional results

Section 1: Chapter 4 results

Table 1: Odds ratios for child participation at different ages: baseline covariates

Covariate	Level	OR (95% CI)		
		< 11 years	11-15 years	16+ years
Sex	Female vs male	1.62 (1.45, 1.80)	2.50 (2.16, 2.89)	3.61 (3.18, 4.09)
Mother's education	O level / lower	1.00	1.00	1.00
	A level	1.47 (1.28, 1.68)	1.74 (1.44, 2.10)	1.66 (1.42, 1.95)
	Degree/higher	1.68 (1.40, 2.01)	1.87 (1.47, 2.40)	2.28 (1.86, 2.80)
Parity	0	1.00	1.00	1.00
	1	0.77 (0.68, 0.88)	0.60 (0.50, 0.72)	0.70 (0.68, 0.88)
	2	0.61 (0.50, 0.74)	0.36 (0.27, 0.47)	0.50 (0.50, 0.74)
	3+	0.46 (0.34, 0.61)	0.23 (0.15, 0.34)	0.38 (0.27, 0.55)
Mother's age (at	<20	1.00	1.00	1.00
birth of index child)	20-24	1.82 (1.23, 2.69)	2.26 (1.32, 3.88)	2.78 (1.23, 2.69)
	25-29	3.01 (2.01, 4.49)	4.07 (2.34, 7.06)	3.01 (2.01, 4.49)
	30-34	4.42 (2.89, 6.77)	6.40 (3.56, 11.50)	4.11 (2.89, 6.77)
	35+	5.03 (3.18, 7.96)	8.56 (4.55, 16.12)	5.75 (3.20, 10.36)
Mother's ethnicity	Non-white vs white	0.47 (0.32, 0.69)	0.50 (0.29, 0.69)	0.69 (0.43, 1.10)
Age at first	<20	1.00	1.00	1.00
pregnancy	20-24	1.31 (1.10, 1.57)	1.39 (1.08, 1.79)	1.22 (0.97, 1.52)
	25-29	1.40 (1.14, 1.71)	1.40 (1.06, 1.85)	1.47 (1.16, 1.88)
	30+	1.42 (1.10, 1.86)	1.43 (0.99, 2.05)	1.55 (1.14, 2.11)
Maternal smoking	Y vs N (in pregnancy)	0.78 (0.66, 0.92)	0.71 (0.57, 0.89)	0.95 (0.79, 1.16)
	Y vs N (ever)	0.86 (0.75, 0.97)	0.65 (0.55, 0.78)	0.66 (0.57, 0.77)
Duration of	Never/<1 month	1.00	1.00	1.00
breastfeeding	1 to <3 months	1.65 (1.39, 1.95)	2.00 (1.58, 2.54)	1.55 (1.27, 1.90)
	3 to <6 months	1.78 (1.51, 2.09)	2.21 (1.77, 2.77)	1.87 (1.55, 2.26)
	6 months+	2.17 (1.89, 2.49)	3.00 (2.48, 3.64)	2.36 (2.01, 2.78)
Married	Yes vs no	1.15 (0.99, 1.34)	1.01 (0.82, 1.24)	1.03 (0.86, 1.23)
Depression score	Per 1 unit increase	0.99 (0.98, 1.00)	0.99 (0.97, 1.00)	0.99 (0.98, 1.01)
Family social class	Manual vs non-manual	0.85 (0.73, 0.99)	0.77 (0.62, 0.95)	0.78 (0.65, 0.94)
Housing tenure	Owned/mortgaged	1.00	1.00	1.00
	Private rented	0.64 (0.50, 0.82)	0.52 (0.37, 0.73)	0.68 (0.51, 0.92)
	Council/HA/other	0.85 (0.70, 1.03)	0.73 (0.56, 0.96)	0.74 (0.58, 0.94)
Number of rooms	Per 1 room increase	1.06 (1.01, 1.12)	1.14 (1.06, 1.21)	1.08 (1.02, 1.14)
Phone in home	Yes vs no/incoming	0.71 (0.56, 0.89)	0.60 (0.43, 0.82)	0.71 (0.53, 0.94)
Car use	No vs yes	0.66 (0.52, 0.84)	0.58 (0.42, 0.81)	0.80 (0.59, 1.08)
Double glazing	None vs full/partial	0.85 (0.76, 0.95)	0.89 (0.76, 1.04)	0.90 (0.79, 1.03)
Financial difficulties	Per 1 unit increase	0.98 (0.96, 1.00)	0.97 (0.95, 0.99)	0.97 (0.95, 0.99)

Section 2: Chapter 5 results

Table 2: Breastfeeding and IQ: results from inverse probability weighting with large weights truncated

Weights truncated at	Breastfeeding duration	Unadjusted	Adjusted
Excluding linked variables			
8	Never / < 1 month	--	--
	1 to <3 months	2.0 (0.4, 3.5)	0.7 (-0.8, 2.2)
	3 to <6 months	5.5 (4.2, 6.7)	2.7 (1.5, 3.9)
	6 months +	8.1 (7.0, 9.1)	3.7 (2.6, 4.8)
6	Never / < 1 month	--	--
	1 to <3 months	2.0 (0.5, 3.5)	0.8 (-0.6, 2.2)
	3 to <6 months	5.4 (4.1, 6.6)	2.7 (1.5, 3.9)
	6 months +	8.0 (7.0, 9.0)	3.7 (2.6, 4.8)
4	Never / < 1 month	--	--
	1 to <3 months	2.0 (0.5, 3.4)	0.8 (-0.5, 2.1)
	3 to <6 months	5.3 (4.0, 6.5)	2.7 (1.5, 3.9)
	6 months +	7.9 (6.8, 8.9)	3.7 (2.6, 4.7)
Including linked variables			
8	Never / < 1 month	--	--
	1 to <3 months	2.5 (0.8, 4.1)	1.6 (0.0, 3.1)
	3 to <6 months	4.9 (3.5, 6.3)	3.0 (1.7, 4.4)
	6 months +	8.1 (6.9, 9.4)	4.4 (3.2, 5.7)
6	Never / < 1 month	--	--
	1 to <3 months	2.3 (0.7, 3.9)	1.3 (-0.2, 2.9)
	3 to <6 months	4.7 (3.3, 6.1)	2.9 (1.6, 4.2)
	6 months +	7.9 (6.7, 9.1)	4.2 (3.0, 5.5)
4	Never / < 1 month	--	--
	1 to <3 months	2.0 (0.4, 3.6)	1.1 (-0.4, 2.6)
	3 to <6 months	4.4 (3.1, 5.8)	2.7 (1.4, 4.0)
	6 months +	7.5 (6.4, 8.7)	4.0 (2.8, 5.2)

Appendix B: Additional results

Table 3: Complete case and MI estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR, generated using a logistic model: OR_{obs} = 2 for every 1 SD increase in IQ

		Complete case		MI	
Scenario	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias
IQ 20% missing	β_4	0.06 (0.034)	-41%	0.08 (0.033)	-23%
Correlation(IQ:KS4)	β_5	0.14 (0.032)	-29%	0.17 (0.030)	-16%
= 0.7	β_6	0.23 (0.025)	-22%	0.26 (0.024)	-13%
IQ 40% missing	β_4	0.04 (0.038)	-55%	0.07 (0.035)	-31%
Correlation(IQ:KS4)	β_5	0.12 (0.035)	-41%	0.15 (0.033)	-24%
= 0.7	β_6	0.20 (0.030)	-33%	0.24 (0.027)	-19%
IQ 60% missing	β_4	0.05 (0.050)	-52%	0.07 (0.042)	-30%
Correlation(IQ:KS4)	β_5	0.12 (0.044)	-40%	0.15 (0.038)	-23%
= 0.7	β_6	0.20 (0.038)	-34%	0.24 (0.033)	-20%
IQ 80% missing	β_4	0.06 (0.078)	-39%	0.08 (0.061)	-23%
Correlation(IQ:KS4)	β_5	0.14 (0.063)	-28%	0.17 (0.052)	-16%
= 0.7	β_6	0.22 (0.055)	-27%	0.25 (0.047)	-15%

Table 4: Complete case and MI estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR: difference in Pr(IQ observed) = 0.05 for 1 SD increase in IQ (20% and 40% missing data)

		Complete case		MI			
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision	FMI
IQ 20% missing Correlation(IQ:KS4) = 0.1	β_4	0.09 (0.034)	-6%	0.09 (0.034)	-5%	0%	24%
	β_5	0.19 (0.031)	-5%	0.19 (0.031)	-4%	0%	21%
	β_6	0.29 (0.025)	-4%	0.29 (0.025)	-4%	0%	22%
IQ 20% missing Correlation(IQ:KS4) = 0.3	β_4	As above		0.09 (0.034)	-5%	2%	23%
	β_5			0.19 (0.031)	-4%	1%	20%
	β_6			0.29 (0.025)	-4%	1%	20%
IQ 20% missing Correlation(IQ:KS4) = 0.5	β_4	As above		0.10 (0.033)	-4%	6%	20%
	β_5			0.19 (0.030)	-4%	4%	17%
	β_6			0.29 (0.025)	-3%	3%	18%
IQ 20% missing Correlation(IQ:KS4) = 0.7	β_4	As above		0.10 (0.032)	-3%	14%	15%
	β_5			0.20 (0.029)	-3%	8%	13%
	β_6			0.29 (0.024)	-2%	8%	13%
IQ 20% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.10 (0.031)	-1%	27%	7%
	β_5			0.20 (0.029)	-1%	16%	6%
	β_6			0.30 (0.024)	-1%	15%	6%
IQ 40% missing Correlation(IQ: KS4) = 0.1	β_4	0.09 (0.039)	-11%	0.09 (0.039)	-11%	0%	44%
	β_5	0.18 (0.035)	-9%	0.18 (0.036)	-9%	1%	41%
	β_6	0.28 (0.030)	-6%	0.28 (0.030)	-6%	0%	42%
IQ 40% missing Correlation(IQ: KS4) = 0.3	β_4	As above		0.09 (0.039)	-10%	3%	43%
	β_5			0.18 (0.035)	-8%	2%	40%
	β_6			0.28 (0.029)	-5%	3%	40%
IQ 40% missing Correlation(IQ: KS4) = 0.5	β_4	As above		0.09 (0.037)	-9%	11%	38%
	β_5			0.19 (0.034)	-7%	8%	35%
	β_6			0.29 (0.028)	-5%	10%	36%
IQ 40% missing Correlation(IQ:KS4) = 0.7	β_4	As above		0.09 (0.035)	-6%	26%	31%
	β_5			0.19 (0.032)	-6%	20%	28%
	β_6			0.29 (0.027)	-3%	22%	29%
IQ 40% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.10 (0.032)	-3%	54%	15%
	β_5			0.19 (0.029)	-3%	46%	13%
	β_6			0.29 (0.024)	-1%	48%	14%

Appendix B: Additional results

Table 5: Complete case and MI estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR: difference in Pr(IQ observed) = 0.05 for 1 SD increase in IQ (60% and 80% missing data)

		Complete case		MI			
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision	FMI
IQ 60% missing Correlation(IQ:KS4) = 0.1	β_4	0.07 (0.049)	-32%	0.07 (0.049)	-32%	-1%	65%
	β_5	0.15 (0.044)	-23%	0.16 (0.042)	-22%	0%	62%
	β_6	0.25 (0.038)	-16%	0.25 (0.038)	-15%	0%	63%
IQ 60% missing Correlation(IQ:KS4) = 0.3	β_4	As above		0.07 (0.048)	-29%	3%	63%
	β_5			0.16 (0.041)	-21%	3%	61%
	β_6			0.26 (0.037)	-14%	5%	61%
IQ 60% missing Correlation(IQ:KS4) = 0.5	β_4	As above		0.08 (0.046)	-25%	14%	59%
	β_5			0.16 (0.041)	-18%	12%	56%
	β_6			0.25 (0.036)	-12%	16%	57%
IQ 60% missing Correlation(IQ:KS4) = 0.7	β_4	As above		0.08 (0.042)	-17%	35%	50%
	β_5			0.17 (0.039)	-13%	30%	47%
	β_6			0.27 (0.032)	-8%	40%	48%
IQ 60% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.09 (0.035)	-6%	93%	29%
	β_5			0.19 (0.031)	-7%	79%	26%
	β_6			0.29 (0.027)	-3%	102%	27%
IQ 80% missing Correlation(IQ:KS4) = 0.1	β_4	-0.09 (0.075)	-187%	-0.08 (0.074)	-185%	0%	85%
	β_5	-0.03 (0.063)	-117%	-0.03 (0.063)	-116%	0%	84%
	β_6	0.06 (0.055)	-79%	0.07 (0.055)	-78%	0%	85%
IQ 80% missing Correlation(IQ:KS4) = 0.3	β_4	As above		-0.07 (0.072)	-173%	6%	84%
	β_5			-0.02 (0.061)	-113%	6%	83%
	β_6			0.08 (0.053)	-73%	4%	84%
IQ 80% missing Correlation(IQ:KS4) = 0.5	β_4	As above		-0.05 (0.068)	-148%	20%	82%
	β_5			0.01 (0.058)	-93%	18%	80%
	β_6			0.11 (0.051)	-63%	16%	82%
IQ 80% missing Correlation(IQ:KS4) = 0.7	β_4	As above		-0.007 (0.061)	-107%	54%	77%
	β_5			0.07 (0.052)	-67%	48%	75%
	β_6			0.16 (0.045)	-45%	48%	76%
IQ 80% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.06 (0.045)	-43%	172%	57%
	β_5			0.15 (0.039)	-27%	157%	54%
	β_6			0.25 (0.034)	-18%	164%	56%

Table 6: Complete case and MI estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR: difference in Pr(IQ observed) = 0.20 for 1 SD increase in IQ (20% and 40% missing data)

		Complete case		MI			
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision	FMI
IQ 20% missing Correlation(IQ:KS4) = 0.1	β_4	0.05 (0.032)	-47%	0.05 (0.032)	-47%	0%	24%
	β_5	0.13 (0.030)	-38%	0.13 (0.030)	-37%	1%	21%
	β_6	0.21 (0.024)	-30%	0.21 (0.024)	-30%	1%	21%
IQ 20% missing Correlation(IQ:KS4) = 0.3	β_4	As above		0.06 (0.032)	-44%	1%	23%
	β_5			0.13 (0.030)	-35%	3%	20%
	β_6			0.22 (0.024)	-28%	2%	20%
IQ 20% missing Correlation(IQ:KS4) = 0.5	β_4	As above		0.06 (0.031)	-38%	4%	21%
	β_5			0.14 (0.029)	-30%	6%	17%
	β_6			0.23 (0.024)	-24%	5%	17%
IQ 20% missing Correlation(IQ:KS4) = 0.7	β_4	As above		0.07 (0.031)	-28%	9%	16%
	β_5			0.16 (0.029)	-22%	12%	13%
	β_6			0.25 (0.023)	-17%	9%	13%
IQ 20% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.09 (0.029)	-12%	18%	7%
	β_5			0.18 (0.028)	-9%	18%	6%
	β_6			0.28 (0.023)	-7%	13%	6%
IQ 40% missing Correlation(IQ: KS4) = 0.1	β_4	0.03 (0.039)	-67%	0.03 (0.039)	-66%	0%	47%
	β_5	0.10 (0.033)	-49%	0.10 (0.033)	-48%	1%	42%
	β_6	0.19 (0.029)	-38%	0.19 (0.029)	-38%	1%	42%
IQ 40% missing Correlation(IQ: KS4) = 0.3	β_4	As above		0.04 (0.038)	-62%	2%	45%
	β_5			0.11 (0.033)	-45%	3%	41%
	β_6			0.19 (0.028)	-35%	2%	41%
IQ 40% missing Correlation(IQ: KS4) = 0.5	β_4	As above		0.05 (0.037)	-53%	7%	41%
	β_5			0.12 (0.032)	-39%	8%	37%
	β_6			0.21 (0.027)	-30%	8%	37%
IQ 40% missing Correlation(IQ:KS4) = 0.7	β_4	As above		0.06 (0.036)	-39%	17%	33%
	β_5			0.14 (0.031)	-29%	17%	30%
	β_6			0.23 (0.026)	-22%	19%	30%
IQ 40% missing Correlation(IQ:KS4) = 0.9	β_4	As above		0.08 (0.033)	-16%	38%	16%
	β_5			0.18 (0.029)	-12%	35%	14%
	β_6			0.27 (0.024)	-9%	39%	14%

Table 7: Complete case and MI estimates of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR: difference in Pr(IQ observed) = 0.20 for 1 SD increase in IQ (60% and 80% missing data)

		Complete case		MI			
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precision	FMI
IQ 60% missing	β_4	-0.02 (0.047)	-125%	-0.02 (0.047)	-124%	0.1%	67%
Correlation(IQ: KS4)	β_5	0.02 (0.041)	-92%	0.02 (0.041)	-90%	0.4%	64%
= 0.1	β_6	0.10 (0.035)	-67%	0.10 (0.035)	-66%	0.4%	64%
IQ 60% missing	β_4	As above		-0.02 (0.046)	-116%	4.2%	66%
Correlation(IQ: KS4)	β_5			0.03 (0.040)	-85%	4.5%	63%
= 0.3	β_6			0.11 (0.034)	-63%	4.3%	63%
IQ 60% missing	β_4	As above		-0.001 (0.044)	-101%	14%	63%
Correlation(IQ:KS4)	β_5			0.05 (0.039)	-74%	13%	59%
= 0.5	β_6			0.14 (0.033)	-54%	13%	59%
IQ 60% missing	β_4	As above		0.03 (0.040)	-74%	34%	54%
Correlation(IQ: KS4)	β_5			0.09 (0.036)	-55%	32%	50%
= 0.7	β_6			0.18 (0.030)	-40%	30%	51%
IQ 60% missing	β_4	As above		0.07 (0.034)	-31%	85%	32%
Correlation(IQ:KS4)	β_5			0.15 (0.031)	-24%	78%	28%
= 0.9	β_6			0.25 (0.026)	-17%	73%	29%
IQ 80% missing	β_4	-0.12 (0.064)	-219%	-0.12 (0.064)	-219%	-0.3%	86%
Correlation(IQ: KS4)	β_5	-0.13 (0.054)	-166%	-0.13 (0.054)	-165%	-0.9%	85%
= 0.1	β_6	-0.08 (0.045)	-125%	-0.07 (0.046)	-125%	-0.3%	85%
IQ 80% missing	β_4	As above		-0.11 (0.063)	-208%	3.2%	86%
Correlation(IQ: KS4)	β_5			-0.11 (0.053)	-157%	2.6%	84%
= 0.3	β_6			-0.06 (0.045)	-119%	2.5%	84%
IQ 80% missing	β_4	As above		-0.08 (0.060)	-184%	13%	84%
Correlation(IQ: KS4)	β_5			-0.08 (0.051)	-139%	11%	82%
= 0.5	β_6			-0.02 (0.043)	-105%	10%	83%
IQ 80% missing	β_4	As above		-0.04 (0.055)	-142%	36%	80%
Correlation(IQ:KS4)	β_5			-0.01 (0.047)	-107%	31%	77%
= 0.7	β_6			0.06 (0.040)	-81%	29%	78%
IQ 80% missing	β_4	As above		0.04 (0.043)	-64%	120%	61%
Correlation(IQ:KS4)	β_5			0.11 (0.038)	-48%	96%	57%
= 0.9	β_6			0.19 (0.033)	-37%	94%	58%

Table 8: Complete case, MI and IPW estimates[1] of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR with an interaction[2] between breastfeeding and IQ with respect to the probability of IQ being observed (20% missing data)

		Complete case		MI				IPW		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion	FMI	Estimate (empirical SE)	% bias	Gain in precis-ion
IQ 20% missing	β_4	0.06 (0.034)	-42%	0.06 (0.034)	-41%	1%	24%	0.06 (0.034)	-42%	0%
Correlation	β_5	0.13 (0.031)	-36%	0.13 (0.031)	-36%	0%	21%	0.13 (0.031)	-36%	0%
(IQ:KS4) = 0.1	β_6	0.21 (0.026)	-31%	0.21 (0.026)	-30%	1%	21%	0.21 (0.026)	-31%	1%
IQ 20% missing	β_4	As above		0.06 (0.034)	-39%	3%	23%	0.06 (0.034)	-40%	-1%
Correlation	β_5			0.13 (0.030)	-33%	2%	20%	0.13 (0.031)	-34%	0%
(IQ: KS4) = 0.3	β_6			0.22 (0.026)	-28%	2%	20%	0.21 (0.026)	-29%	-1%
IQ 60% missing	β_4	As above		0.07 (0.033)	-33%	8%	24%	0.06 (0.035)	-35%	-1%
Correlation	β_5			0.14 (0.030)	-28%	6%	21%	0.14 (0.031)	-31%	0%
(IQ: KS4) = 0.5	β_6			0.23 (0.025)	-24%	5%	21%	0.22 (0.026)	-27%	-1%
IQ 20% missing	β_4	As above		0.08 (0.032)	-24%	15%	15%	0.07 (0.035)	-28%	-2%
Correlation	β_5			0.16 (0.029)	-20%	12%	13%	0.15 (0.031)	-25%	-1%
(IQ: KS4) = 0.7	β_6			0.25 (0.025)	-17%	11%	13%	0.23 (0.026)	-23%	-2%
IQ 20% missing	β_4	As above		0.09 (0.030)	-9%	27%	7%	0.08 (0.035)	-17%	-5%
Correlation	β_5			0.18 (0.028)	-8%	20%	5%	0.17 (0.031)	-17%	-2%
(IQ: KS4) = 0.9	β_6			0.28 (0.024)	-7%	19%	6%	0.25 (0.026)	-17%	-4%

1. FIML results not included as they were very similar to the MI results
2. Difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when exposure=0 (no breastfeeding); change in difference in Pr(IQ observed) for each 1 SD increase in IQ = -0.025 for each increase in breastfeeding category (Factor 4 in scenarios)

Table 9: Complete case, MI and IPW estimates[1] of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR with an interaction[2] between breastfeeding and IQ with respect to the probability of IQ being observed (40% missing data)

		Complete case		MI				IPW		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion	FMI	Estimate (empirical SE)	% bias	Gain in precis-ion
IQ 20% missing	β_4	0.03 (0.040)	-66%	0.03 (0.040)	-65%	-1%	44%	0.03 (0.040)	-66%	-1%
Correlation(IQ:	β_5	0.10 (0.036)	-51%	0.10 (0.036)	-51%	0%	41%	0.10 (0.036)	-52%	-1%
KS4) = 0.1	β_6	0.17 (0.030)	-44%	0.17 (0.030)	-44%	-1%	41%	0.17 (0.030)	-45%	-1%
IQ 20% missing	β_4	As above		0.04 (0.039)	-61%	3%	42%	0.03 (0.040)	-65%	-2%
Correlation(IQ:	β_5			0.11 (0.035)	-47%	4%	40%	0.10 (0.036)	-51%	-2%
KS4) = 0.3	β_6			0.18 (0.030)	-41%	2%	40%	0.17 (0.030)	-44%	-1%
IQ 60% missing	β_4	As above		0.05 (0.038)	-52%	10%	38%	0.04 (0.040)	-63%	-2%
Correlation(IQ:	β_5			0.12 (0.034)	-40%	10%	36%	0.10 (0.036)	-49%	-2%
KS4) = 0.5	β_6			0.20 (0.029)	-35%	9%	36%	0.17 (0.031)	-43%	-1%
IQ 20% missing	β_4	As above		0.06 (0.036)	-37%	23%	31%	0.03 (0.040)	-59%	-3%
Correlation(IQ:	β_5			0.14 (0.032)	-29%	24%	28%	0.11 (0.036)	-46%	-3%
KS4) = 0.7	β_6			0.23 (0.027)	-25%	22%	29%	0.18 (0.031)	-40%	-2%
IQ 20% missing	β_4	As above		0.08 (0.032)	-15%	49%	15%	0.05 (0.041)	-53%	-5%
Correlation(IQ:	β_5			0.18 (0.029)	-11%	50%	13%	0.12 (0.036)	-41%	-4%
KS4) = 0.9	β_6			0.28 (0.025)	-10%	51%	14%	0.19 (0.031)	-37%	-4%

1. FIML results not included as they were very similar to the MI results
2. Difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when exposure=0 (no breastfeeding); change in difference in Pr(IQ observed) for each 1 SD increase in IQ = -0.025 for each increase in breastfeeding category (Factor 4 in scenarios)

Table 10: Complete case, MI and IPW estimates[1] of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR with an interaction[2] between breastfeeding and IQ with respect to the probability of IQ being observed (60% missing data)

		Complete case		MI				IPW		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion	FMI	Estimate (empirical SE)	% bias	Gain in precis-ion
IQ 20% missing	β_4	-0.03 (0.048)	-132%	-0.03 (0.048)	-131%	-1%	63%	-0.04 (0.048)	-138%	-2%
Correlation(IQ:	β_5	0.003 (0.043)	-99%	0.005 (0.043)	-98%	-1%	63%	-0.003 (0.043)	-102%	-1%
KS4) = 0.1	β_6	0.06 (0.037)	-79%	0.06 (0.037)	-79%	0%	63%	0.06 (0.037)	-81%	0%
IQ 20% missing	β_4	As above		-0.02 (0.048)	-123%	2%	63%	-0.04 (0.048)	-138%	-3%
Correlation(IQ:	β_5			0.02 (0.043)	-91%	2%	61%	-0.002 (0.043)	-101%	-1%
KS4) = 0.3	β_6			0.08 (0.037)	-73%	4%	61%	0.06 (0.037)	-81%	0%
IQ 60% missing	β_4	As above		-0.005 (0.046)	-105%	11%	59%	-0.04 (0.048)	-137%	-3%
Correlation(IQ:	β_5			0.04 (0.041)	-78%	11%	57%	0.0003 (0.043)	-100%	-2%
KS4) = 0.5	β_6			0.11 (0.035)	-63%	15%	57%	0.06 (0.037)	-80%	0%
IQ 20% missing	β_4	As above		0.02 (0.042)	-76%	30%	51%	-0.03 (0.048)	-135%	-4%
Correlation(IQ:	β_5			0.09 (0.038)	-56%	30%	48%	0.004 (0.043)	-98%	-3%
KS4) = 0.7	β_6			0.17 (0.032)	-45%	38%	49%	0.06 (0.037)	-78%	-1%
IQ 20% missing	β_4	As above		0.07 (0.035)	-31%	85%	29%	-0.03 (0.048)	-133%	-5%
Correlation(IQ:	β_5			0.16 (0.032)	-22%	79%	27%	0.008 (0.043)	-96%	-5%
KS4) = 0.9	β_6			0.26 (0.026)	-18%	99%	28%	0.07 (0.037)	-77%	-3%

1. FIML results not included as they were very similar to the MI results
2. Difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when exposure=0 (no breastfeeding); change in difference in Pr(IQ observed) for each 1 SD increase in IQ = -0.025 for each increase in breastfeeding category (Factor 4 in scenarios)

Table 11: Complete case, MI and IPW estimates[1] of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR with an interaction[2] between breastfeeding and IQ with respect to the probability of IQ being observed (80% missing data)

		Complete case		MI				IPW		
Scenario (Factors 1 & 2)	Estimand	Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias	Gain in precis-ion	FMI	Estimate (empirical SE)	% bias	Gain in precis-ion
IQ 20% missing	β_4	-0.23 (0.070)	-330%	-0.23 (0.070)	-328%	1%	85%	-0.27 (0.069)	-370%	2%
Correlation(IQ:	β_5	-0.28 (0.062)	-242%	-0.28 (0.062)	-240%	-1%	84%	-0.34 (0.062)	-269%	-2%
KS4) = 0.1	β_6	-0.25 (0.054)	-184%	-0.25 (0.054)	-183%	0%	84%	-0.30 (0.054)	-200%	-1%
IQ 20% missing	β_4	As above		-0.21 (0.068)	-307%	6%	84%	-0.27 (0.069)	-372%	1%
Correlation(IQ:	β_5			-0.25 (0.061)	-225%	4%	83%	-0.34 (0.062)	-270%	-2%
KS4) = 0.3	β_6			-0.21 (0.052)	-171%	5%	83%	-0.30 (0.054)	-201%	-2%
IQ 60% missing	β_4	As above		-0.16 (0.064)	-263%	18%	82%	-0.28 (0.069)	-376%	2%
Correlation(IQ:	β_5			-0.19 (0.058)	-193%	14%	80%	-0.35 (0.062)	-272%	-2%
KS4) = 0.5	β_6			-0.14 (0.050)	-147%	16%	81%	-0.31 (0.054)	-203%	-3%
IQ 20% missing	β_4	As above		-0.09 (0.058)	-190%	47%	77%	-0.28 (0.069)	-382%	1%
Correlation(IQ:	β_5			-0.08 (0.052)	-140%	41%	74%	-0.35 (0.063)	-277%	-3%
KS4) = 0.7	β_6			-0.02 (0.045)	-106%	43%	76%	-0.32 (0.055)	-206%	-5%
IQ 20% missing	β_4	As above		0.02 (0.044)	-78%	147%	57%	-0.29 (0.070)	-391%	1%
Correlation(IQ:	β_5			0.09 (0.040)	-57%	135%	53%	-0.37 (0.063)	-285%	-6%
KS4) = 0.9	β_6			0.17 (0.034)	-43%	147%	56%	-0.34 (0.056)	-212%	-9%

1. FIML results not included as they were very similar to the MI results
2. Difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when exposure=0 (no breastfeeding); change in difference in Pr(IQ observed) for each 1 SD increase in IQ = -0.025 for each increase in breastfeeding category (Factor 4 in scenarios)

Table 12: Complete case and MI results of β_4, β_5 and β_6 (true values 0.1, 0.2, 0.3) for IQ MNAR: difference in Pr(IQ observed) = 0.10 for 1 SD increase in IQ when linked attainment associated with sex and mother's education in addition to IQ

Scenario	Estimand	Complete case[1]		MI	
		Estimate (empirical SE)	% bias	Estimate (empirical SE)	% bias
IQ 20% missing	β_4	0.08 (0.034)	-17%	0.09 (0.033)	-12%
Correlation(IQ:KS4)	β_5	0.17 (0.030)	-14%	0.18 (0.029)	-8%
= 0.7	β_6	0.26 (0.025)	-12%	0.28 (0.027)	-8%
IQ 40% missing	β_4	0.07 (0.041)	-29%	0.08 (0.037)	-15%
Correlation(IQ:KS4)	β_5	0.16 (0.035)	-20%	0.18 (0.032)	-11%
= 0.7	β_6	0.26 (0.029)	-15%	0.28 (0.026)	-8%
IQ 60% missing	β_4	0.03 (0.049)	-74%	0.06 (0.043)	-44%
Correlation(IQ:KS4)	β_5	0.10 (0.043)	-49%	0.14 (0.037)	-30%
= 0.7	β_6	0.19 (0.037)	-36%	0.24 (0.032)	-21%
IQ 80% missing	β_4	-0.14 (0.068)	-237%	-0.05 (0.059)	-147%
Correlation(IQ:KS4)	β_5	-0.13 (0.062)	-165%	-0.0002 (0.053)	-100%
= 0.7	β_6	-0.05 (0.052)	-116%	0.09 (0.045)	-71%

1. These results are the same as those shown in Tables 5-11 to 5-14; they are given here for completeness.

Section 3: Chapter 7 results

Table 13: Teenage smoking and educational attainment: results from IPW with large weights truncated (linked variables included in IPW)

Exposure variable	Weights truncated at:	Binary outcome		Continuous outcome	
		Crude	Adjusted	Crude	Adjusted
Ever-smoked	10	3.27 (2.53, 4.23)	2.75 (1.95, 3.86)	-55 (-63, -47)	-29 (-34, -24)
	8	3.27 (2.53, 4.23)	As above	-55 (-63, -47)	As above
	6	3.26 (2.52, 4.22)	As above	-55 (-63, -47)	As above
	4	3.23 (2.50, 4.18)	As above	-53 (-61, -46)	-29 (-33, -24)
Frequency of smoking	10	7.58 (5.07, 11.35)	8.33 (5.00, 13.88)	-103 (-126, -79)	-67 (-82, -52)
	8	7.39 (4.96, 11.01)	8.26 (4.95, 13.78)	-100 (-121, -78)	-65 (-80, -51)
	6	7.05 (4.76, 10.43)	8.22 (4.92, 13.73)	-95 (-114, -75)	-63 (-76, -50)
	4	6.59 (4.49, 9.68)	8.20 (4.91, 13.72)	-87 (-103, -71)	-60 (-71, -48)
Cotinine	10	5.21 (3.36, 8.09)	5.51 (3.03, 10.01)	-90 (-112, -67)	-56 (-72, -41)
	8	5.12 (3.31, 7.91)	5.39 (2.97, 9.77)	-88 (-109, -66)	-55 (-70, -40)
	6	4.87 (3.17, 7.48)	5.24 (2.90, 9.49)	-83 (-102, -63)	-53 (-67, -39)
	4	4.54 (2.97, 6.93)	4.14 (2.63, 6.53)	-76 (-94, -58)	-50 (-63, -37)

Section 4: Chapter 8 results

Table 14: Effect of smoking at 15 on educational attainment: comparison of methods to take account of misclassification: estimates (log odds ratio and regression coefficient) in four additional simulated datasets with true smoking MCAR

Dataset	Analysis method	N	Did not obtain 5+ more A*-C grades	KS4 attainment score
2	Naive analyses			
	True smoking, complete	100,000	1.617 (1.573, 1.661)	-44.3 (-45.8, -42.9)
	True smoking, observed	60,036	1.633 (1.575, 1.690)	-44.0 (-45.9, -42.1)
	Self-reported smoking	100,000	1.762 (1.711, 1.812)	-52.6 (-54.2, -51.0)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.622 (1.568, 1.676)	-45.9 (-47.6, -44.1)
	Multiple imputation – 1[1]	100,000	1.619 (1.567, 1.671)	-44.7 (-46.3, -43.0)
	Multiple imputation – 2[2]	100,000	1.620 (1.568, 1.672)	N/A
	Bayesian analysis	100,000	1.622 (1.569, 1.674)	-45.1 (-46.9, -43.5)
3	Naive analyses			
	True smoking, complete	100,000	1.625 (1.581, 1.669)	-43.9 (-45.4, -42.5)
	True smoking, observed	59,873	1.623 (1.566, 1.678)	-43.6 (-45.5, -41.8)
	Self-reported smoking	100,000	1.758 (1.707, 1.808)	-52.2 (-53.8, -50.6)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.623 (1.570, 1.677)	-45.0 (-46.8, -43.3)
	Multiple imputation – 1[1]	100,000	1.623 (1.573, 1.674)	-43.5 (-45.1, -41.8)
	Multiple imputation – 2[1]	100,000	1.622 (1.571, 1.673)	N/A
	Bayesian analysis	100,000	1.623 (1.573, 1.674)	-44.2 (-45.8, -42.5)
4	Naive analyses			
	True smoking, complete	100,000	1.595 (1.551, 1.638)	-43.6 (-45.0, -42.2)
	True smoking, observed	59,945	1.593 (1.537, 1.649)	-44.4 (-46.2, -42.5)
	Self-reported smoking	100,000	1.723 (1.673, 1.773)	-51.1 (-52.6, -49.5)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.581 (1.527, 1.634)	-44.7 (-46.4, -43.0)
	Multiple imputation – 1[1]	100,000	1.581 (1.531, 1.632)	-43.8 (-45.4, -42.2)
	Multiple imputation – 2[1]	100,000	1.581 (1.531, 1.631)	N/A
	Bayesian analysis	100,000	1.580 (1.532, 1.633)	-43.8 (-45.5, -42.2)
5	Naive analyses			
	True smoking, complete	100,000	1.620 (1.576, 1.664)	-43.9 (-45.3, -42.4)
	True smoking, observed	59,985	1.603 (1.547, 1.660)	-43.5 (-45.4, -41.6)
	Self-reported smoking	100,000	1.769 (1.718, 1.820)	-52.6 (-54.2, -51.0)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.635 (1.579, 1.690)	-45.5 (-47.2, -43.8)
	Multiple imputation – 1[1]	100,000	1.634 (1.584, 1.685)	-44.2 (-45.9, -42.5)
	Multiple imputation – 2[1]	100,000	1.633 (1.582, 1.684)	N/A
	Bayesian analysis	100,000	1.636 (1.585, 1.685)	-44.6 (-46.3, -42.9)

1. Multiple imputation – 1: continuous outcome included in imputations when estimating effect of smoking on binary outcome; multiple imputation – 2: continuous outcome excluded from same.

Table 15: Effect of smoking at 15 on educational attainment: comparison of methods to take account of misclassification: estimates (log odds ratio and regression coefficient) in four additional simulated datasets with true smoking MNAR

Dataset	Analysis method	N	Did not obtain 5+ more A*-C grades	KS4 attainment score
2	Naive analyses			
	True smoking, complete	100,000	1.643 (1.599, 1.687)	-44.5 (-45.9, -43.0)
	True smoking, observed	59,543	1.566 (1.506, 1.627)	-42.4 (-44.4, -40.3)
	Self-reported smoking	100,000	1.749 (1.698, 1.800)	-52.4 (-54.0, -50.8)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.624 (1.569, 1.678)	-45.8 (-47.6, -44.0)
	Multiple imputation – 1[1]	100,000	1.625 (1.570, 1.680)	-44.3 (-46.1, -42.5)
	Multiple imputation – 2[1]	100,000	1.625 (1.572, 1.677)	N/A
	Bayesian analysis	100,000	1.626 (1.572, 1.676)	-45.5 (-47.3, -43.7)
3	Naive analyses			
	True smoking, complete	100,000	1.618 (1.575, 1.662)	-44.2 (-45.7, -42.8)
	True smoking, observed	59,176	1.601 (1.540, 1.661)	-42.6 (-44.7, -40.5)
	Self-reported smoking	100,000	1.782 (1.732, 1.833)	-53.6 (-55.2, -52.0)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.616 (1.562, 1.672)	-45.6 (-47.2, -43.8)
	Multiple imputation – 1[1]	100,000	1.612 (1.559, 1.665)	-43.2 (-45.0, -41.5)
	Multiple imputation – 2[1]	100,000	1.616 (1.563, 1.669)	N/A
	Bayesian analysis	100,000	1.619 (1.572, 1.670)	-45.0 (-46.8, -43.2)
4	Naive analyses			
	True smoking, complete	100,000	1.608 (1.564, 1.651)	-44.3 (-45.8, -42.9)
	True smoking, observed	59,515	1.565 (1.504, 1.625)	-41.5 (-43.6, -39.4)
	Self-reported smoking	100,000	1.731 (1.681, 1.781)	-51.8 (-53.4, -50.2)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.589 (1.535, 1.643)	-45.1 (-46.8, -43.4)
	Multiple imputation – 1[1]	100,000	1.587 (1.533, 1.640)	-43.4 (-45.1, -41.6)
	Multiple imputation – 2[1]	100,000	1.590 (1.536, 1.644)	N/A
	Bayesian analysis	100,000	1.590 (1.538, 1.641)	-44.7 (-46.5, -43.0)
5	Naive analyses			
	True smoking, complete	100,000	1.635 (1.591, 1.678)	-44.2 (-45.8, -42.8)
	True smoking, observed	59,553	1.588 (1.528, 1.648)	-42.5 (-44.5, -40.4)
	Self-reported smoking	100,000	1.770 (1.720, 1.820)	-52.4 (-54.0, -50.8)
	Methods used to correct for misclassification			
	Probabilistic bias analysis	100,000	1.613 (1.559, 1.667)	-45.3 (-47.0, -43.6)
	Multiple imputation – 1[1]	100,000	1.613 (1.559, 1.668)	-43.8 (-45.6, -42.0)
	Multiple imputation – 2[1]	100,000	1.611 (1.555, 1.668)	N/A
	Bayesian analysis	100,000	1.614 (1.567, 1.663)	-44.9 (-46.7, -43.2)

1. Multiple imputation – 1: continuous outcome included in imputations when estimating effect of smoking on binary outcome; multiple imputation – 2: continuous outcome excluded from same.